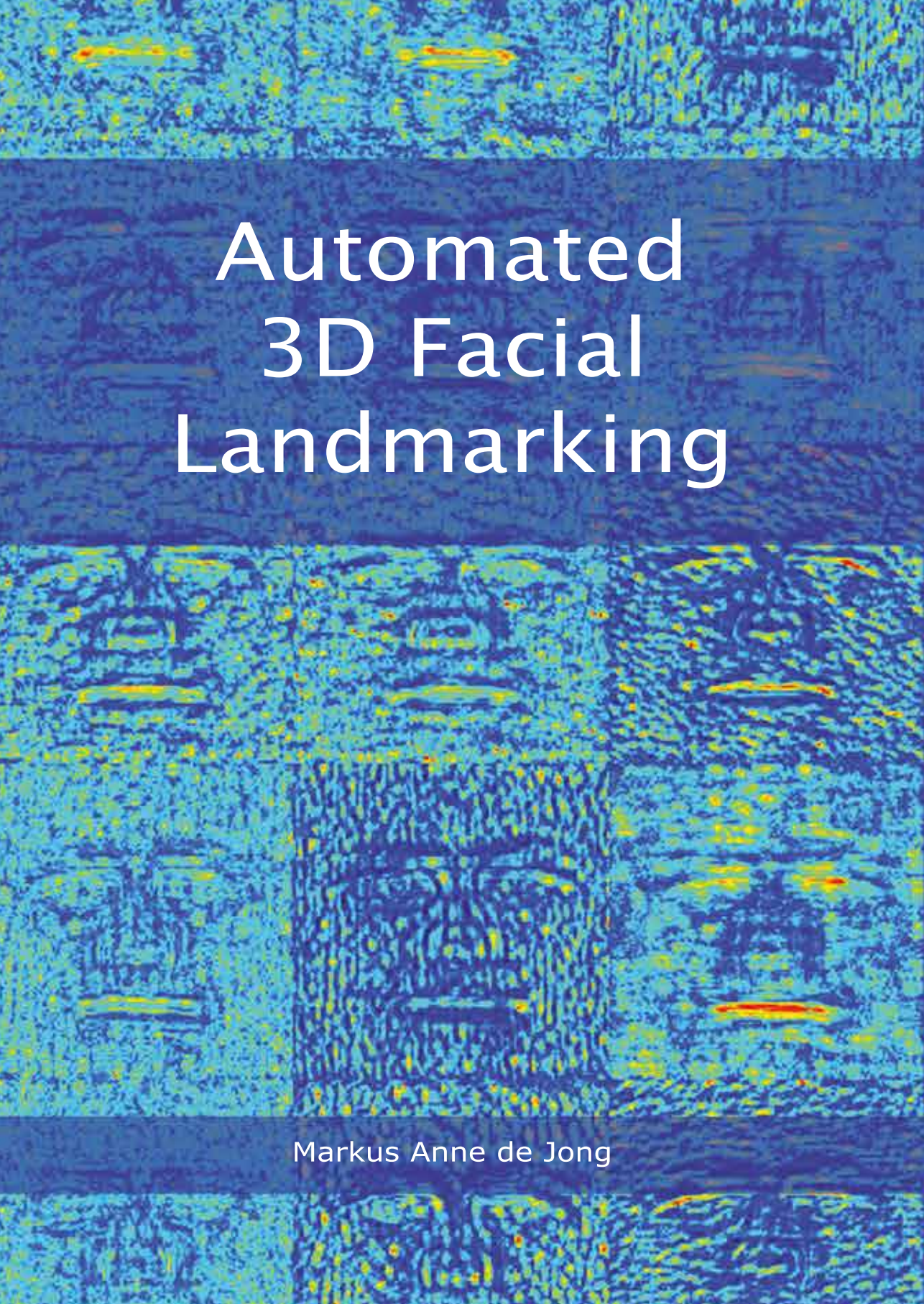

Automated
3D Facial
Landmarking
Markus Anne de Jong

Stellingen

behorende bij het proefschrift
Automated 3D facial landmarking

Stellingen related to the dissertation

chapter 2, first paper: **1: 2D Gabor wavelets are suitable to solve 3D computer vision problems**

chapter 3, ensemble paper: **2: Sometimes the forest can be better seen for the trees: more and smart features improve landmarking**

chapter 4, clinical paper: **3: Our algorithm is ready for clinical use**

chapter 5, skull landmarking: **4: A good and useful algorithm works on different types of data**

chapter 6, symmetry: **5: An algorithm based on our landmarks can be used to perform symmetry analysis**

*Stellingen **not** related to the dissertation*

6: "You must always remember that the products of your mind can be used by other people either for good or for evil, and that you have a responsibility that they be used for good." Dean Llewellen M. K. Boelter

7: Artificial intelligence has a racial bias problem

8: Deep learning is promising but can be beaten for small training samples

9: We will at some time be able to predict a face only from a DNA sample - or will we not?

10: We will be able to automatically judge a face and tell its owner's gender, ethnic origins and medical baggage and ultimately success in life

Vrije stelling

We should all put a 3D scan of our faces online - it's a visible trait.

Markus de Jong
October 4th, 2019

Automated 3D Facial Landmarking

Geautomatiseerde 3D-gezichtslabeling

Thesis

To obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the rector magnificus
Prof.dr. R.C.M.E. Engels
and in accordance with the decision of the Doctorate Board
The public defense shall be held on
October 4th at 13:30

by

Markus Anne de Jong
born in Workum

Erasmus University Rotterdam

DOCTORAL COMMITTEE

Promotors:

Prof.dr. E.B. Wolvius

Prof.dr. M. Kayser

Co-Promotors:

Dr. S. Böhringer

Dr. M.J. Koudstaal

Other members:

Dr. H. Bosch

Dr.ir. F. van der Heijden

Prof.dr. M. Reinders

Paranimphs:

Lieuwe van der Meer

Arthur Melissen

Table of contents

Manuscripts that form the basis of this thesis

de Jong, Markus A., Andreas Wollstein, Clifford Ruff, David Dunaway, Pirro Hysi, Tim Spector, Fan Liu, Wiro Niessen, Maarten J. Koudstaal, Manfred Kayser, Eppo B. Wolvius, Stefan Böhringer. "An automatic 3d facial landmarking algorithm using 2d gabor wavelets." IEEE Transactions on Image Processing 25, no. 2 (2016): 580-588.

de Jong, Markus A., Pirro Hysi, Tim Spector, Wiro Niessen, Maarten J. Koudstaal, Eppo B. Wolvius, Manfred Kayser, and Stefan Böhringer. "Ensemble landmarking of 3D facial surface scans." Scientific reports 8, no. 1 (2018): 12.

de Jong, Markus A., Atilla Gül, Jan Pieter de Gijt, Maarten J. Koudstaal, Manfred Kayser, Eppo B. Wolvius, and Stefan Böhringer. "Automated human skull landmarking with 2D Gabor wavelets." Physics in medicine and biology (2018).

Atilla Gül, **de Jong, Markus A.**, Jan Pieter de Gijt, Eppo B. Wolvius, Manfred Kayser, Stefan Böhringer, and Maarten J. Koudstaal. "Three-dimensional soft tissue effects of mandibular midline distraction and surgically assisted rapid maxillary expansion: an automatic stereophotogrammetry landmarking analysis." International Journal of Oral and Maxillofacial Surgery (2018).

de Jong, Markus A., Eppo B. Wolvius, Pirro Hysi, Tim Spector, Manfred Kayser, and Stefan Böhringer. "Automated asymmetry estimation of facial 3D scans using guaranteed symmetrical correspondence". *Under review*

1

General introduction

1.1 Introduction

Our facial features are one of the most identifiable traits that we possess and play a large part in our daily social interactions. Their importance may perhaps best be objectively illustrated by research that indicates that the visual system of our brain is specialized towards the processing of faces, where each face is reconstructed using combinations of dedicated sets of neurons.[8]

There is much to learn from any face we are confronted with, even if we are not always consciously doing so. For example, we can either recognize a person from memory or determine someone as a stranger. We can classify a person by identifying the person's gender, age and ethnicity. Facial expressions convey information about our current emotions and support our social interactions. The shape of the mouth supports our speech recognition. We can appraise fitness and health, which we may translate in terms as normalcy and attractiveness that we use in, for example, mate selection. Because these facial aspects are of such great interest to us humans, the face is subject of study in a broad set of scientific research areas ranging from psychology to genetics, and from medicine to security and forensics.

Often, medical research investigates non-normalcy. For example, facial asymmetry is clinically relevant in relation to movement-related problems of the jaw, while dysmorphism (*i.e.* abnormal facial appearance in general) is used in syndrome diagnosis.

However, before any such research can take place, the facial research data in the face must first be somehow extracted from it. Such facial research data exists in many types, such as shape or size.

Another way to objectively quantify faces is with 3D landmarks: coordinates of interest that reside on the surface of the face. It must also be noted that our goals differ from facial recognition. Relative landmark positions that are sometimes used for those purposes are insufficient to us, as we will discuss later in this section. Consequently, we intend our individual landmarks to be as accurate as possible and to let them carry anatomical information on their own.

In their unprocessed form, the informative value of 3D facial photographs is nothing more than that of 3D mugshots: only usable for casual visual inspection. More interesting is overlaying several 3D faces over each other for comparison, however doing this manually takes time. 3D landmarks can be used to assist in overlaying 3D photographs. For example, isolating interesting landmarks such as the eye and mouth corners becomes easier by focusing on those landmarks over a photograph set. For small data sets, manual placement of the landmarks is not a problem. However, many research topics have available much larger data sets or even rely on large sets. Genetics, for example, often requires thousands of labeled faces.

Until a few years ago, extracting the useful data and measurements from facial data was a manual endeavor in which people sat down, visually inspected large amounts of images and labeled each one of them by hand. Now, with the advances in computers and software, we are starting to automate these kinds of boring and repetitive and labor-intensive tasks. But how can you make a computer recognize a landmark location, a face, or how make it see anything at all? This question is researched in a sub-field of computer science called computer vision.

Recently, several high-profile examples that make extensive use of computer vision have become (close to) reality. Examples are automated driving and augmented reality, illustrated in Figure 1.1a and 1.1b. In the medical field, computer vision algorithms are already partnering with clinicians by helping them comb through large amounts of medical imagery to look for signs of cancer.[16] For faces specifically, the savvy social media user will know the popular facial enhancement filters on social media. These filters recognize your face captured in full motion video and automatically overlay the image with entertaining elements such as dog ears and party hats at the correct locations. Other filters are able to swap faces between persons on screen and unlocking your phone by presenting your face to the phone's camera is another much-used example.

Judging by these examples, one could consider the processing of visual data by computers solved. However, the requirements for our algorithm differ in more than one aspect from these existing examples, which will be further explained in the next sections.

The remainder of this introduction is structured as follows: first we introduce the computer vision field and some of its techniques that we will use in our algorithm, all the wile converging towards facial data. Then, we explore the scientific background behind our landmarking algorithm, such as the way we approach our data and machine learning techniques, a kind of artificial intelligence. Finally, a full overview of the process behind this thesis is given and we take a look at its methodology and applications in the remaining chapters.

(a) Automated driving (b) Augmented reality in liver surgery

Figure 1.1: Recent examples of computer vision applications.

1.2 Computer Vision

1.2.1 Introduction

The act of seeing is a feat that humans can do instantly and effortlessly. Perhaps against the expectations of laypeople, 'seeing' is an act with which machines struggle greatly. Even in the early days of artificial intelligence, the general notion was that computers are worse at cognitive functions such as planning and better at perceptive functions. An interesting anecdote from the early days of computer vision perhaps illustrates this misconception best and goes as follows: in 1966, Marvin Minsky at MIT asked his undergraduate to "spend the summer linking a camera to a computer and getting the computer to describe what it saw".[19] Computer scientist now know that this is far more difficult than it seems as only today, some 50 years later, we can find instances where the computer indeed appears to 'see' effortlessly.

Illustrated by the fact that so many years have passed until we have reached the stage in which a computer may 'understand' what he 'sees' well enough to support technology such as autonomous cars, the road to the current state of the art was long and difficult indeed. Even still, recent computer vision technology is still limited to specific sets of circumstances. When, for instance, recognizable road markings disappear, the automated cars will screech to a halt, and when mobile phone filter algorithms are presented with a non-frontal face, they often fail.

Part of the difficulty with computer vision lies in the fact that it needs to solve an *inverse problem*. We are given a visual input and are to recover or reconstruct certain unknowns from incomplete information. There are almost infinite possible solutions and we must select the best one based on, for example, what we know about the physics involved with lighting and by applying probability calculations. Computer vision algorithms must be robust as visual input can vary greatly and often contains a lot of visual 'noise' such as shadows or occlusions. Another great challenge (and one that we do delve into in this thesis) is how easily persons connect meaning to what they witness. In order for a computer to 'interpret' a scene, it must first be taught its meaning bit by bit (or rather 'pixel by pixel').

Besides recent impressive advances as automated driving, mobile phone filters and augmented reality, computer vision plays a role in many other already established techniques. Some early examples include optical character recognition (OCR) to scan and recognize printed texts and handwritten postal addresses. Others such examples are quality assurance by machine inspection of products on a conveyor belt, medical imaging (MRI, CT) and automatically stitching together panorama photographs. Photogrammetry involves the reconstruction of a 3D model from multiple 2D sources and is the technique used to create our 3D facial images. This technique is also applied in other areas such as popular online 3D mapping software, *e.g.* Google Maps. Other examples include motion capture in the film and game industry.

We will now have a look at the computer vision basics and challenges and will work our way towards the kind of methods we apply in our project.

1.2.2 Making computers see a mug

As said, for persons, object recognition is instantaneous. We can rely on millions of years of evolution that has given us a most impressive visual processing system stretching from our eyes to our visual cortex in the brain. Of course, for computers, this skill had to be (re-)constructed from the ground up. In this section, we will attempt to illustrate the computer vision struggle with a thought experiment in which we try to make a computer recognize a coffee mug in any situation, just like a (healthy) person is able to.

When we look at a coffee mug on the table in front of us, we clearly see a whole and distinct shape and are able to recognize it. But how can we make a computer recognize a coffee mug? For a computer, the basic input it receives from its camera is a 2D rectangular matrix of unrelated, colored pixels. This has no meaning, each pixel is as important as the other. We need to find a way to recognize the group of pixels that we, as humans, recognize as a coffee mug. Let's call our mug M.

We need to compare the input image, the 'test image', against some kind of example: a reference picture of a coffee mug. We call this reference picture a 'training sample' that we use to 'train' our coffee mug recognition algorithm. Using our training sample, we could now, somehow, calculate the difference between the input image and our training image.

A simple method is to subtract the pixel values that represent the colors and brightness in each pixel of our training sample from the input image, and sum all the individual pixel differences up. The closer the distance d between input and training sample is to zero, the more equal the two images are and the better the match is:

$$d = Test\_image - Train\_image \tag{1.1}$$

We can now define the distance on which our system will detect a mug, $d\_threshold$. Equations 1.2 and 1.3 define the two matching outcomes.

$$\text{if } d \text{ larger than } d\_threshold, \text{ then } Test\_image = M (\text{a match}) \tag{1.2}$$

$$\text{if } d \text{ smaller than } d\_threshold, \text{ then } Test\_image \neq M (\text{NOT a match}) \tag{1.3}$$

We can lower $d\_threshold$ to add a little bit of flexibility so it does not need to be an exact 1:1. However, it is important to note that while adding this flexibility, we will then also open the door to

false matches.

But what if our input image has a much smaller representation of a mug? Or what if the mug is oriented differently, or has a different color? And what about the background that will also influence the pixel difference score? This method will only work if the training and test images have the same composition. Otherwise, this simple method of overlaying input and training images 1:1 and subtracting them does not work.

A first solution is to isolate the mug in our training examples. This way, we only need to compare a small area against each area in the test image by 'scanning', perform calculation 1.4 only on small parts ($Scanning\_part$) and we can forget about the background. But the problem of mug flexibility needs another solution. A first one could to use a huge number of training samples of different mugs, colors, shapes and sizes, and that also includes all of their orientations and positions. Creating such a list reference images not only seems like a lot of effort, but this would also create a problem when our (*e.g.* particular mug was not included the training set. Will it still be 'seen'? Also, how long will it take to compare the input image against each of the samples this huge training set and repeating calculation 1.4?

Instead, perhaps a more sensible and more efficient idea is to generalize towards some sort of universal concept of a coffee mug that fits many different types of coffee mugs at the same time. This way, there is no need to keep a near-infinite collection of training images. How would such a 'coffee mug concept' look like? Initially, the coffee mug concept should probably include a cylindrical shape and a handle. Then, we will to look for cylinder shapes in close proximity of a handle shape. We now can use this concept to generate many variations from a single concept we can use to scan each part of the test image for coffee mug candidates and perform calculation 1.4 for the parts.

Taking this idea even further, we could subdivide the mug concept into many different, smaller parts, or features, such as corners and edges that, when combined, constitute the mug. For each small feature, we also use different sizes and orientations. This way, we only have to compare small parts to each other which will speed the process up:

$$d = \text{sum all}(Scanning\_part - Train\_features) \tag{1.4}$$

We may now even detect mugs that are partially obscured, as long as enough smaller parts are visible and by setting $d\_threshold$ accordingly.

Now we can quickly scan the input image for our limited set of small coffee mug features. And when we find a configuration of such parts close together, this might indicate the presence of a coffee mug. At first glance, our hypothetical mug detector seems finished, but we are still not there. Of course, we must find a mathematical way to determine if the mug parts are located where they should be. And what if the coffee mug is missing a handle? Will the parts still be recognized if they are in a shadow or have unusual colors or patterns? How many parts are 'enough' to recognize a mug (remember $d\_threshold$)? What about other objects or parts of the scenery that also are cylindrical shaped and have handles, such as tea cups? And can we make our algorithm fast enough for full-motion video?

Even though many questions are left open (especially the question of attaching 'meaning' to what the computer sees), this thought experiment on recognizing a mug has hopefully illustrated that computer

vision requires a very high level of flexibility on many levels. One must keep in mind low level problems (*e.g.* defining the right mug features) and high(er) level problems (*e.g.* categorization and 'meaning'), as well as efficiency considerations like memory optimization and process running times.

On a final note, we here hypothesized a model with a task hierarchy that is constructed from bottom to top, from individual feature to final detection. For completeness, it is important to mention that there are other methods that do not require such a pipeline, nor the defining of particular features. The latest methods can be so-called data-driven and can rely on neural networks to quickly classify mug examples based only on labeled training images.[17] However, as will become clear, such methods do not meet the requirements for our project.

1.2.3 Computer vision and faces

Much computer vision research in relation to faces has been focused on several separate tasks, such as facial detection ("is there a face visible?"), (real-time) identification ("who is this person?") and identity verification ("is this person who he/she claim to be?").

Often, these tasks are the focus in security applications. For example, a program may detect and decide to only record video when a face is present with a security camera. A more lighthearted example of face detection are the social media overlay filters mentioned earlier.

An example of facial identification are facial search engines that are now in use in some states of the US. These search engines have already lead to the arrest of wanted suspects who had started their new lives in another state, only to find their picture of their driver's license automatically picked out of a database of 120 million people.[1]

An example of identity verification unlocking your phone with your face. A local example is an experiment currently being conducted at the Amsterdam Airport Schiphol in the Netherlands, in which facial data on the passport is compared with the image from a camera for customs automation (Figure 1.3a).

Another, perhaps worrisome, development in computer vision in relation to facial data is the digitally re-enacting of faces of high-profile persons, live and in full motion video, and to make them say things they in fact never said, also known as "deep fake".[20]

Such facial recognition algorithms typically rely on the extraction of features or templates, such as eyes and nose, or (pseudo-)landmarks and use distances and compositions in a comparison.

There are several important differences in what these algorithms have to offer and what is required for our project. Firstly, our goal is not facial recognition and does not involve complete facial features or pseudo-landmarks. Instead, our goal is the accurate localization of individual and true anatomical landmarks. Secondly, we intend to use rich, highly detailed 3D facial data that is relatively unexplored in the field and for which no (open source) method or algorithm is available. Therefore, 3D data has its own unique opportunities and challenges. Thirdly, the requirements are such that a large amount of flexibility is required. The facial algorithms that are used in the examples above are rigid in that they still require hundreds of manual training examples to learn a new landmark. We should have the ability

[1] http://www.washingtonpost.com/business/technology/state-photo-id-databases-become-troves-for-police/2013/06/16/6f014bd4-ced5-11e2-8845-d970ccb04497_story.html

to quickly change landmark sets during our studies, which goes hand in hand with small training sets. Finally, we want to combine all these points with maximum accuracy by using a proven algorithm.

1.2.4 Beyond pixels

Instead of looking at the level of individual pixels (like in our mug example), a more advanced method also used in facial recognition is using *Gabor Wavelets*.

Invented by Dennis Gabor (1900 – 1979) [7], Gabor Wavelets tend to mimic the workings of the human visual system in that they form a layered deconstruction of a visual scene. Each wavelet has an orientation and size and, when applied to a given location in the image, results in a response. This response is a number that measures how much the neighborhood of that location resembles the wavelet. In this way, we separate an image into features each being the response of a different wavelet with a different orientation. By combining all the detected features, we can reconstruct an image. This reconstruction process is illustrated in Figure 1.2.

The difference between pixels and Gabor Wavelet responses is that wavelet responses have additional interpretation, namely whether a given number corresponds to a larger or smaller wavelet, or which orientation was used. For example, if fine image structures are of interest, only responses of small wavelets have to be analyzed. The same principle is used when an image is rendered over a slow internet connection that gradually shows more detail. Responses corresponding to coarse features are transmitted first and allow to reconstruct broader contours. As more features become available finer and finer features can be shown until the full image is reconstructed (technically, a discrete Fourier transform is used in the web example, which is very similar to a wavelet analysis).

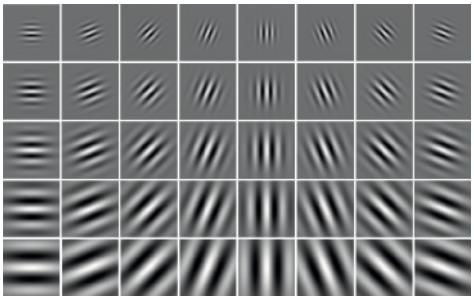
1.3 Landmarking

1.3.1 Introduction

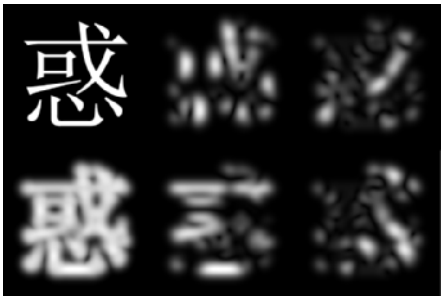
Landmark registration of the human face is an important prerequisite for many epidemiological and clinical applications.[1–4,9–12,14,21] Such studies are concerned with characterizing this trait in terms of heritability,[11] genetic association,[3,12] or syndrome classification.[2,4,9,10,21] Such studies often rely on a specific set of landmarks that are of interest.[4,18] Also, the image acquisition process varies between studies.[4,12,18] An automatic landmarking algorithm should therefore be flexible enough to deal with varying image raw material, changing sets of landmarks, and smaller sets of training data.

Recently, 3D facial data had been used in epidemiological[14,15] and earlier in clinical studies.[9,10] In all but one of these studies only a limited set of landmarks were placed manually. The study that does employ automatic landmarking, does so with strong heuristic components with limited flexibility.[15] In light of this previous work, we aim to develop an algorithm for 3D facial image registration meeting our aims on flexibility and training complexity.

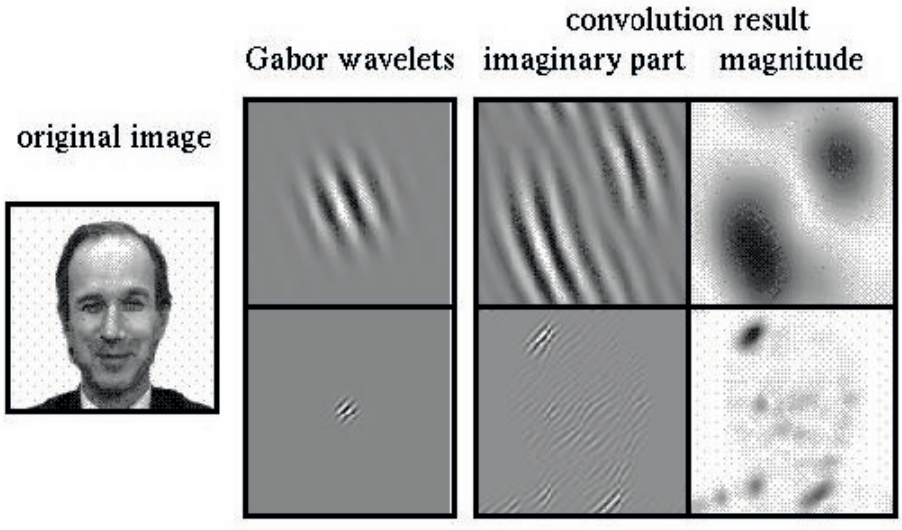
Our approach is to work with 2D projections of 3D surface data and to employ well-studied 2D landmarking algorithms on that transformed data. In this process, we keep all the information about the original surface data. The face-specific components of our algorithm lie in a pre-processing step -



(a) Example Gabor filter set with different orientations and sizes

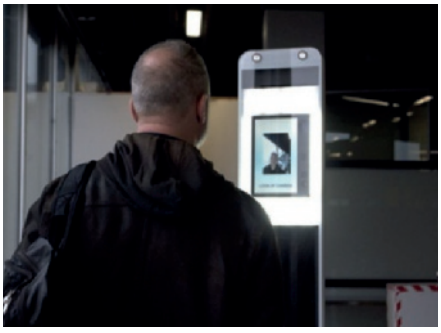


(b) Example decomposition of a Chinese character

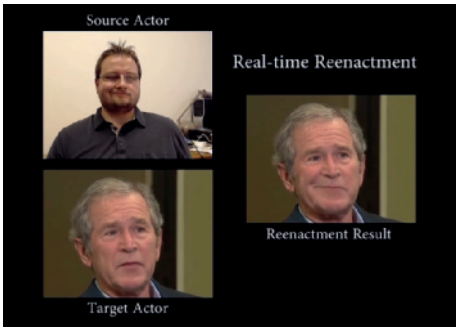


(c) Example decomposition of a face

Figure 1.2: Examples of Gabor filter decompositions.



(a) Identity verification at Schiphol airport, Amsterdam



(b) "Deep fake": Capture and re-enactment of faces in video[20]

Figure 1.3: Recent examples of computer vision applications for facial data.

defining the region of interest - and the projection method. For the landmarking, we choose a Gabor wavelet-based procedure.[22] As stated before, Gabor wavelet-based procedures are well-studied and have the advantage of performing well with few training examples. This contrasts with, for example, active shape models (ASMs) that are *e.g.* used in social media and which need up to thousands of training samples for accurate registration.[6] One important aspect is the richness of 3D data when compared to 2D-data. We want to use this information to the maximum by adding 3D information to our registration algorithm in a generic way, *i.e.* by presenting it as 2D data.

1.3.2 Smarter landmarkers

During our project, we experienced differences in performance for different types of 2D information extracted from the face. Ideally, we only want to use the best performing information and ignore the remainder as this will achieve the most accurate results. In theory, we could perform this selection manually. However, this would be time consuming and inflexible as the set of landmarks are large and the types of 2D information are many. Luckily, machine learning techniques in the form of ensemble methods are suited to automatically select the best features for each landmark from a large set of different features.

Ensemble methods in machine learning can be described as automatically making a combination of multiple learning algorithms to reach a better prediction than any of the individual algorithms. These learning algorithms are allowed to be many different approaches, the only requirement is that they lead to the same type of result, in our case the coordinates of the landmark.

For our landmarking algorithm, we can describe each 2D feature as connected to a separate learning algorithm: a *landmarker*. In the early version of our algorithm, we calculated the final coordinate of a landmark by taking the average of all of the landmarkers' resulting coordinates. This way, 2D features that give good information for a certain landmark are averaged with some that may show poor and erroneous results. By using machine learning to automatically select the best performing set of 2D features for each landmark we can optimize landmarking results.

One way to achieve this optimization is to create an experiment in which we create a very large set of unique combinations of 2D features, and average the total of all those set results. The idea is that the most stable, well-performing predictors automatically surface as these will be in the majority. This ensemble method is called *Bagging* (**b**ootstrap **agg**regating).

Another approach is to create a new learning algorithm that uses the outputs of all the landmarking algorithms. This new combiner algorithm makes a final prediction using all the predictions of the other landmarkers. This method is usually called *stacking* or stacked generalization.

1.4 Automatic landmarking for epidemiological and clinical research

The Erasmus MC is currently involved in two longitudinal studies that are collecting a large variety of data on thousands of cohort subjects from cohorts, including digital 3D images of the subject's face created with photogrammetry and their complete DNA profile. These studies are *ERGO* and *Generation R*.

The ERGO (Erasmus Rotterdam Gezondheid Onderzoek) is a longitudinal or prospective cohort study of more than 15.000 subjects of 40 years and older from the Rotterdam Ommoord area. Its focus lies on aging-related health issues[2].

The Generation R Study is another prospective cohort study, but one that focuses on fetal life until young adulthood. The study is designed to identify early environmental and genetic causes of normal and abnormal growth, development and health from fetal life until young adulthood[3]. Subjects are invited to return every 3 years.

On a smaller scale, 3D images are recorded for clinical purposes of patients that undergo maxillo-facial surgery at the Erasmus MC. In contrast with the cohort studies, these subjects suffer from bone growth syndromes that have resulted into facial abnormalities. Time series that include pre- and post-surgery moments are also recorded.

A first goal of facial 3D data analysis is clinical research: to make 3D facial data accessible to surgery planning and surgery outcome evaluation. On their own, the individual 3D images taken pre-operatively that may be used for surgery planning. Sets of images taken from before and after surgeries allow for pre- and post-surgery comparisons can assist surgeons in their work as well, for example to give insight into the effects of a surgery. A chapter on a clinical application is included in this thesis.

Another clinical aspect of advanced use of the 3D data is the creation of growth curves. An example question that may be answered with such growth curves is to pinpoint the optimal age to undergo syndrome-related facial surgery. If the growth curves show no change in facial non-normality over the years, one could, for example, conclude it is not necessary to perform surgery at a young age and that it is possible to wait for a more suitable moment.

A second goal is genetic research. In the past decade, genetic association study techniques have become commonplace. These studies allow to assess the association of genome-wide set of genetic variants with certain complex traits such as diseases, but also traits such as skin [13], hair and eye color [5]. Such a a study is known as a genome-wide association study, or GWAS. Now, due to the availability of large cohorts that contain both complete genetic information and 3D facial data, using GWASes to investigate the genetic origins of the facial shape has become possible.

Although current research is still ongoing, a hypothetical application of using DNA and facial shape lies in the forensic science field. However, due to the underlying complexity and environmental effects, simply predicting facial features based on a DNA sample and printing a mugshot for the police cannot be realized. The results would be too ambiguous and would only be detrimental to an investigation.

[2] **http://www.erasmus-epidemiology.nl/research/ergo.htm**.
[3] **https://www.generationr.nl/**.

A more realistic approach would be to turn the process around and exclude a person by comparing a given image of a face with the predicted results from a certain unknown DNA profile.

The research at Rotterdam is part of a consortium-wide GWAS study where 3D data from different sources and countries are combined. For this purpose, circa 3000 facial models from the ERGO set were automatically landmarked for 21 landmarks. Comparative studies based on the Generation R cohort are still in the planning phase.

1.5 This thesis

The focus of this dissertation is the creation and application of software for the automatic landmarking of large sets of digital 3D facial models that can label the 3D locations of points of interest (*i.e.* landmarks) for clinical and genetic-forensic research purposes.

Chapter 2 introduces the automatic landmarking algorithm and forms the foundation on which all other work is built. In this chapter, a novel method is presented that involves loss-less map projections of 3D facial images that convert the information to 2D. From this map projection, we extract many modalities of 2D information and use this as input for an established automatic 2D landmarking algorithm that locates the landmarks. The 2D landmarks and projected face are then reverted back to 3D coordinates. The results are validated with a leave-one-out study design in which a 3D face is taken from a set and the algorithm is trained with the remainder of the faces. The trained algorithm is then applied to the face that was taken out. This process is repeated for all the faces in the set. As such, we are able to perform an independent experiment for each of the 3D faces. To further to illustrate the algorithm's validity, a complex heritability-based study of identical twins is performed. Here, we use the known genetic information of identical twins to compare landmarking performance.

Chapter 3 covers the enhancements made to the existing automated algorithm by using ensemble methods. Experiments are carried out with different machine learning methods aimed towards the automatic selection of the best performing 2D information for each landmark. We also make use of the natural grouping of landmarks to predict where, for example, in what position and orientation a group of eye corners are most likely be found. Again, we validate our results using a twin heritability study.

Chapter 4 describes a clinical application of our 3D landmarking algorithm in a pre- and post-operation comparison of Erasmus MC patients of the maxillo-facial surgery department. Two types of facial surgery are compared that take place in the lower and upper jaw. To investigate the changes that occur as a result of the surgeries, the landmarks are located and a statistical investigation is performed on the inter-landmark distances.

Chapter 5 investigates the development and use of an altered version our our facial landmarking algorithm that is aimed towards human 3D skulls. Besides being among the first to explore automated landmarking to 3D CT-scans, this paper also illustrates the flexibility of our algorithm. A leave-one-out study design and a relevant practical application of skull super-imposition are used to illustrate its effectiveness.

Chapter 6 shows an investigation into the determination of facial symmetry. An important step needed for the comparison of left and right halves of the face is the registration step. This registration

is supported by a set of automatically located landmarks. To investigate the accuracy of our registration method, we apply controlled deformations to a standardized 3D face and subject it to our registration algorithm and compare the registration with the original.

References

[1] Brunilda Balliu et al. "Classification and Visualization Based on Derived Image Features: Application to Genetic Syndromes". In: *PloS one* 9.11 (2014), e109033. URL: http://dx.plos.org/10.1371/journal.pone.0109033 (visited on 05/28/2015).

[2] Stefan Boehringer et al. "Automated syndrome detection in a set of clinical facial photographs". In: *American Journal of Medical Genetics Part A* 155.9 (2011), pp. 2161–2169. ISSN: 1552-4833.

[3] Stefan Boehringer et al. "Genetic determination of human facial morphology: links between cleft-lips and normal variation". In: *European Journal of Human Genetics* 19.11 (2011), pp. 1192–1197. ISSN: 1018-4813.

[4] Stefan Boehringer et al. "Syndrome identification based on 2D analysis software". In: *European Journal of Human Genetics: EJHG* 14.10 (Oct. 2006), pp. 1082–9. ISSN: 1018-4813. DOI: 5201673. URL: http://www.ncbi.nlm.nih.gov/pubmed/16773127 (visited on 11/13/2008).

[5] Lakshmi Chaitanya et al. "The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation". In: *Forensic Science International: Genetics* (2018).

[6] Timothy F. Cootes et al. "Active shape models-their training and application". In: *Computer vision and image understanding* 61.1 (1995), pp. 38–59. URL: http://www.sciencedirect.com/science/article/pii/S1077314285710041 (visited on 06/05/2014).

[7] Dennis Gabor. "Theory of communication. Part 1: The analysis of information". In: *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering* 93.26 (1946), pp. 429–441.

[8] Avniel Singh Ghuman et al. "Dynamic encoding of face information in the human fusiform gyrus". In: *Nature communications* 5 (2014), p. 5672.

[9] Peter Hammond et al. "Discriminating Power of Localized Three-Dimensional Facial Morphology". In: *The American Journal of Human Genetics* 77.6 (Dec. 2005), pp. 999–1010. ISSN: 0002-9297. DOI: 10.1086/498396. URL: http://www.sciencedirect.com/science/article/pii/S0002929707633849 (visited on 02/05/2014).

[10] Peter Hammond et al. "Fine-grained facial phenotype–genotype analysis in Wolf–Hirschhorn syndrome". en. In: *European Journal of Human Genetics* 20.1 (Jan. 2012), pp. 33–40. ISSN: 1018-4813. DOI: 10.1038/ejhg.2011.135. URL: http://www.nature.com/ejhg/journal/v20/n1/full/ejhg2011135a.html (visited on 02/05/2014).

[11] L. A. P. Kohn. "The Role of Genetics in Craniofacial Morphology and Growth". In: *Annual Review of Anthropology* 20 (Jan. 1991), pp. 261–278. ISSN: 0084-6570. URL: http://www.jstor.org/stable/2155802 (visited on 07/16/2014).

[12] Fan Liu et al. "A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans". In: *PLoS Genetics* 8.9 (2012), e1002932. ISSN: 1553-7404.

[13] Fan Liu et al. "Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up". In: *Human genetics* 134.8 (2015), pp. 823–835.

[14] Lavinia Paternoster et al. "Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in {\emph PAX3} Associated with Nasion Position". In: *The American Journal of Human Genetics* 90.3 (2012), pp. 478–485. URL: http://www.sciencedirect.com/science/article/pii/S000292971200002X (visited on 06/05/2014).

[15] Shouneng Peng et al. "Detecting Genetic Association of Common Human Facial Morphological Variation Using High Density 3D Image Registration". In: *PLoS computational biology* 9.12 (2013), e1003375. URL: http://dx.plos.org/10.1371/journal.pcbi.1003375.g004 (visited on 06/05/2014).

[16] Alexander Rakhlin et al. "Deep convolutional neural networks for breast cancer histology image analysis". In: *International Conference Image Analysis and Recognition*. Springer. 2018, pp. 737–744.

[17] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[18] Harald J. Schneider et al. "A Novel Approach to the Detection of Acromegaly: Accuracy of Diagnosis by Automatic Face Classification". In: *J Clin Endocrinol Metab* (Apr. 2011), jc.2011–0237. DOI: 10.1210/jc.2011-0237 . URL: http://jcem.endojournals.org/cgi/content/abstract/jc.2011-0237v1 (visited on 05/06/2011).

[19] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[20] Justus Thies et al. "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[21] Tobias Vollmar et al. "Impact of geometry and viewing angle on classification accuracy of 2D based analysis of dysmorphic faces". In: *European Journal of Medical Genetics* 51.1 (2008), pp. 44–53. ISSN: 1769-7212. DOI: S1769-7212(07)00104-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/18054308 (visited on 11/13/2008).

[22] Laurenz Wiskott et al. "Face recognition by elastic bunch graph matching". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19.7 (1997), pp. 775–779. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=598235 (visited on 06/05/2014).

2

Automatic landmarking with 2D Gabor wavelets

Markus A. de Jong
Andreas Wollstein
Clifford Ruff
David Dunaway
Pirro Hysi
Tim Spector
Fan Liu
Wiro Niessen
Maarten J. Koudstaal
Manfred Kayser
Eppo B. Wolvius
Stefan Böhringer

Abstract

In this paper we present a novel approach to automatic 3D facial landmarking using 2D Gabor wavelets. Our algorithm considers the face to be a surface and uses map projections to derive 2D features from raw data. Information extracted includes texture, relief map and transformations thereof. We extend an established 2D landmarking method for simultaneous evaluation of this data. The method is validated by performing landmarking experiments on two data sets using 21 landmarks and compared to an active shape model implementation. On average, landmarking errors were estimated to be 1-2mm for salient landmarks in the eyes, mouth and nose. The active shape model performed at 2-3mm of landmarking error. A second validation using heritability in related individuals shows that automatic landmarking is on par with manual landmarking for some landmarks. Our algorithm can be trained in 30 minutes to automatically landmark 3D facial data sets of any size, and allows for fast and robust landmarking of 3D faces. This mostly non-heuristic implementation makes it flexible to be used on heterogeneous input data and has applications for medical surface 3D data analysis.

2.1 Introduction

Landmark registration of the human face is an important pre-requisite in many epidemiological and clinical applications.[2–5,7–10,12,15] Such studies are concerned with characterizing this trait in terms of heritability,[9] genetic association,[4,10] or the delineation of conditions with characteristic facial morphology.[3,5,7,8,15] In such studies, often a specific set of landmarks is of interest.[5,14] Also, the image acquisition process varies between studies.[5,10,14] An automatic image acquisition process should therefore be flexible to both deal with varying image raw material, changing sets of landmarks, and smaller sets of training data.

Recently, 3D surface scans have been employed in epidemiological[12,13] and earlier in clinical studies.[7,8] In all but one of these studies landmarking was performed manually on a limited set of landmarks.[13] The latter study employs automatic landmarking with strong heuristic components, thereby limiting flexibility.[13] In the present study, we aim to develop an algorithm for 3D facial image registration meeting our aims on flexibility and training complexity.

Our approach is to work with 2D-projections of 3D-surface data and to employ well-studied 2D-algorithms on that transformed data. In this process, we retain complete information about the original surface data. The face-specific components of our algorithm lie in a pre-processing step - defining the region of interest - and the projection method. For the landmarking, we here choose a Gabor-wavelet based procedure.[16] Gabor-wavelet based procedures are well-studied and have the advantage of working well with few training examples. This contrasts, for example, with active shape models (ASMs) which need up to thousands of training samples for accurate registration.[6] One important aspect is the increased richness of 3D data as compared to 2D-data which we exploit by adding 3D information to our registration algorithm in a generic way, *i.e.* by presenting it as 2D data. We evaluate several 3D information components with respect to their impact on registration accuracy by which we evaluate the flexibility of our approach concerning changing landmarking needs. We also evaluate performance in

the context of a heritability study and compare the proposed method to an ASM approach.

The paper is organized as follows: In section two, we initially present an overview of the algorithm and subsequently describe the steps of the algorithm in detail. We also describe evaluation methodology here. In section three, we present an evaluation scenario using cross-validation methodology, perform accuracy evaluation for 3D components on two data sets, and evaluate the contribution of the 3D components. Comparison with an ASM is described in this section. In section four, we evaluate accuracy on unseen data using twin correlation. We conclude with a discussion of limitations and potentials of our approach.

2.2 The Automatic 3D Landmarking Algorithm

2.2.1 Overview

Our algorithm consists of the following steps: First, a region of interest is extracted from the frontal face. Second, a map projection of this face transforms the 3D data set into a 3D relief map. Third, from 3D relief map a 2D image is generated. Fourth, this image is subjected to a 2D landmarking method. In this paper, we make use of the trained Elastic Bunch Graph Matching method (EBGM).[16] Finally, registered 2D landmarks are mapped back into 3D, inverting the projection.

The input of the algorithm is 3D image files of a participant's face obtained with commercial photogrammetry systems for faces called *3dMDface*[1] that creates a 3D surface model without any further user interaction. The output of the system is a triangulation of the 3D surface and a 2D texture for which each point uniquely corresponds to a point in one of the triangles. All data analyzed in this study were recorded with structured light-based triangulation and were exported into the Wavefront *.obj* file format. This format uses vertex indexing that keeps the relations between vertices intact from beginning to end of the algorithm. Projection only uses the vertices of the model (point cloud) and as all transformations are continuous, triangulation is retained throughout the algorithm.

2.2.2 Region of Interest

Landmarking algorithms in general strongly benefit from data preprocessing to remove noise and standardize the input. We use a face-specific, heuristic, preprocessing step to achieve higher landmarking accuracy. For the data sets used in this study, the 3D frontal face models generally include the top of the shoulders, neck and the face itself, but not the back nor any other areas outside the view of the camera system (see image 2.1A).

In order to properly select the region of interest (ROI), *i.e.* the frontal, upright face, the raw 3D facial images have to be rotated upright. This is accomplished via a two-stage ellipsoid fitting process. In the first stage we compensate for unwanted rolling and pitching of the face by freely fitting an ellipsoid to the point cloud using a least square fitting procedure, which minimizes the length of surface normals connecting the ellipsoid and the 3D model. The point cloud is then transformed and rotated upright by using the rotation parameters of the fitted ellipsoid (see Figure 2.1B). In a second stage, we

Figure 2.1: The ellipsoid fitting and map projection process of the human face. A: original point cloud, B: ellipsoid fitting, C: 3D map projection based on ellipsoid

fit a standardized ellipsoid (*i.e.* equal axes ratio) to the upright model to match the shape of the (front of the) head using least squares. This ellipsoid is used for the map projection.

2.2.3 Map Projection

Using the ellipsoid obtained in the previous step, the texture of the 3D face model is projected onto the surface of the ellipsoid and a Mercator map projection is applied to the ellipsoid, using a standard, iterative algorithm. The conversion from Cartesian (x,y,z) to ellipsoidal coordinates latitude ϕ, longitude λ, height h is accomplished as follows:

Longitude λ is given by:

$$\lambda = \arctan \frac{y}{x} \tag{2.1}$$

The iteration procedure for calculating latitude ϕ and height h is as follows:

The initial value is given by:

$$\varphi_0 = \arctan \left[\frac{z}{(1-e^2)p} \right] \tag{2.2}$$

with

$$p = \sqrt{x^2 + y^2} \tag{2.3}$$

Here, e denotes Euler's number. Improved values of ϕ and h are computed by iterating between the following equations until convergence as defined by a preset precision:

$$N_i = \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi_{i-1}}} \tag{2.4}$$

$$h_i = \frac{p}{\cos \varphi_{i-1}} - N_i \tag{2.5}$$

$$\varphi_i = \arctan \left[\frac{z}{\left(1 - e^2 \frac{N_i}{N_i + h_i}\right) p} \right] \tag{2.6}$$

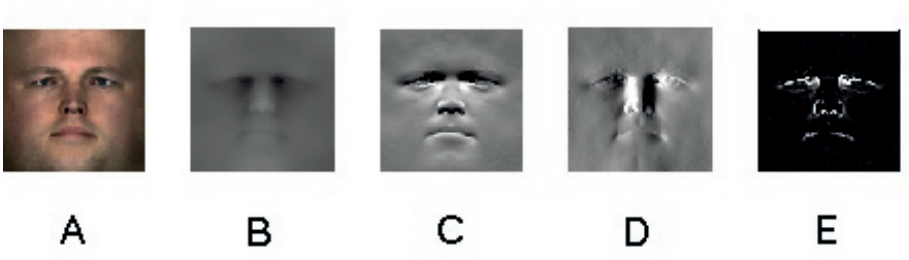


Figure 2.2: The 5 feature layers generated from a human face image based on the map projection. A: photographic, B: heightmap, C: derivative of heightmap with respect to X-axis, D: Y-axis, E: Laplacian of Gaussian of heightmap

For each point of the map, the height of the 3D model above or below the surface of the ellipsoid is stored as a 3D relief map (see image 2.1C). Together with the parameters of the ellipsoid this transformation is therefore one-to-one, *i.e.* we can reconstruct the original 3D model from this data. As a final standardization step, all resulting images are centered at the highest elevation of the relief map, corresponding to the nose tip.

2.2.4 Image Feature Layer Generation

To maximally exploit the available 3D information, several transformations are applied to the 3D relief map to create new features that are potentially useful for subsequent automated landmarking. In total, five feature layers are constructed as follows. First, the texture of the 3D model is rendered orthographically using the 3D editing software Blender under full brightness conditions, i.e. without artificial shadows or specular reflections. Second, the relief map (heightmap) is constructed. The final 3 feature layers are derivatives with respect to the y-axis (layer 3), derivatives with respect to the x-axis (layer 4), and the Laplacian of Gaussian (layer 5). Figure 2.2 contains examples of the 5 generated feature layers.

2.2.5 Training and Landmarking

We applied the EBGM algorithm to the set of feature images. EBGM is described elsewhere in detail.[16] In short, a maximum correlation template search is performed between a set of example images and the image to be landmarked. The features used are Gabor-Wavelet transforms centered at landmarks. If such a landmark is located at pixel $\vec{x} = (x, y)$, the wavelet transform is described by:

$$\begin{aligned} J(\vec{x}) &= (J_1(\vec{x}), ..., J_{40}(\vec{x})), \\ J_j(\vec{x}) &= \int I(\vec{z})\psi_j(\vec{z} - \vec{x})d^2\vec{z}, \end{aligned} \tag{2.7}$$

where $I : \mathbb{R}^2 \to [0, 1]$ represents a grayscale image and ψ_j is a family of Gabor kernels:

$$\psi_j(\vec{x}) = \frac{||\vec{k}_j||^2}{\sigma^2} \exp(-\frac{||\vec{k}_j||^2 \cdot ||\vec{x}||^2}{2\sigma^2}) \left[\exp(i\vec{k}_j^T \cdot \vec{x}) - \exp(-\frac{\sigma^2}{2})\right]. \quad (2.8)$$

Here, $\vec{k}_j$ is the wave vector controlling direction and frequency and σ^2 is the parameter controlling the surface area of the Gabor wavelet.

The search is performed per landmark with global constraints based on deformations of an average graph. We extended the EBMG to run on all feature layers simultaneously, *i.e.* Gabor-Wavelet coefficients were extracted from all layers using the same Gabor-Wavelets and resulting coefficients were combined into a single vector (jet) of 40 kernels (5 wave frequencies $\times$ 8 wave orientations) for each landmark and feature combination. Coefficients per feature layer were standardized to unit variance prior to integration into the jet. In this paper, we used a graph of 21 landmarks, corresponding to anatomical features described in Table 2.1 and illustrated in Figure 2.3.

In the training phase, the EBGM algorithm was trained using 30 images with the 21 landmarks being placed manually. Training landmarks in the layers had one-by-one correspondence, such that training could be carried out on the texture layer and reused for the remaining layers. The mean graph of the training set is used as a starting position for automatic landmarking.

Finally, 2D landmark pixel coordinates are mapped back to the nearest points in the relief map model and subsequently mapped back to 3D landmark coordinates in the 3D cloud. This is achieved by using the inverses of the mapping performed or, more efficiently, using vertex indexing.

2.2.6 Feature importance

Feature importance can be evaluated by considering subsets of features. We evaluate performance of each feature together with all other features (with feature, 'W') and also performance by leaving out each feature one by one (without feature, 'W/O'). If accuracy of a landmark for feature left out drops with respect to all features combined, the feature is essential for accurately labeling that landmark. If accuracy increases in the same comparison, other features contain more information about the given landmark and the feature is not essential. If accuracy stays the same, redundant information is present across features.

2.2.7 Comparison with Active Shape Models

We used a recently published implementation of an ASM (Stasm, version 4.1.0) for comparison.[11] ASMs build statistical models for describing landmark location probability, by establishing correspondence between landmarks in training data, and performing principal component analysis to describe the main variations in shape. As such they tend to require big training data to reliably estimate principal components. The above implementation includes a face model based on ca. 3000 frontal face photographs which was used for our analysis. In order to obtain optimal 2D images for the analysis, perspective projection of frontally aligned 3D models were generated. The same hand-labeled images that were used for evaluating the EBGM on the twinsUK data set were used. Hand-labeled landmarks were transformed through the same projection, defining the ground-truth for this analysis. 18 of Stasm

landmarks coincided with landmarks from our set of 21 landmarks and could be used for comparison (Table 2.1).

2.2.8 Heritability

One way to assess the accuracy of the facial landmarking is to consider faces of related individuals. Informally, heritability can be defined as the proportion of variance explained by "relatedness". Errors of landmarking procedures should add noise to landmark coordinates thereby lowering heritability estimates. Heritability can therefore be used to judge landmarking accuracy independently of comparing automatically with manually derived landmarks the latter of which being subject to rater errors.

The *TwinsUK* data set contains both monozygotic and dizygotic twins. Under the assumption of a polygenic model it is possible to estimate heritability of a trait from such a sample.[17] We used the following random effects model:

$$Y_i = \beta_0 + \beta_1 \text{age}_i + \sigma_2 u_i + \epsilon_i, \tag{2.9}$$

where Y_i is a distance between two landmarks for individual i, age_i is the age, u_i is the random effect and ϵ_i is the residual error. The vector $u = (u_1, ..., u_N)^T$ is assumed to be distributed according to a multivariate normal distribution with mean 0 and covariance matrix Σ: $u \sim MVN(0, \Sigma)$. The entries of Σ are given by the coefficients of relationship between pairs of individuals, *i.e.* $\Sigma_{ij} = 1$ if the pair (i, j) is a monozygotic twin pair, $\Sigma_{ij} = \frac{1}{2}$ for dizygotic twins, $\Sigma_{ii} = 1$, and zero otherwise. u_i is scaled by σ_2 which measures the variance explained by the polygenic effect. ϵ_i is assumed to be an independent residual error normally distributed as $\epsilon_i \sim N(0, \sigma_1^2)$. Heritability can then be estimated by:

$$\hat{h}^2 = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} \tag{2.10}$$

2.3 Experiments

We tested our algorithm using several analyses. First, we performed a 30-fold leave one out experiment using random samples from two different cohorts. The ground truth consists of a single manual labeling of the entire dataset. Second, we evaluate performance of inidivual features. Third, we evaluated the accuracy of the ASM for comparison using the same sub-sample from one of the data sets (TwinsUK) using the same gold standard. The face model provided by the implementation was used for the analysis. Forth, we performed a heritability analysis that can be applied even in the absence of a gold standard.

2.3.1 Data sets

The two datasets used for assessing performance of the presented algorithm are *TwinsUK* and *MeIn3D*.

The *TwinsUK* cohort consists of individuals of full European descent. The cohort consists of volunteers drawn from the general British population, unaware of any 3D studies scientific interests at the time of enrollment and gave fully informed consent under a protocol reviewed by the St. Thomas' Hospital Local Research Ethics Committee. Reference: PMID 23088889.

The *MeIn3D* cohort consists of adults of various ethnicities. The *MeIn3D* research project has been approved by the NHS Research Ethics Committee and is a collaboration between Great Ormond Street Hospital, University College London Hospital and the Eastman Dental Institute.

When comparing the resolution of both datasets, the *TwinsUK* dataset is relatively less detailed with models with ca. 1.5×10^5 points and textures with resolution of ca. 2000x1000 pixels, while the *MeIn3D* dataset contains models with ca. 7.5×10^5 points and textures with resolution of ca. 5000x4000 pixels. 3D images of both data sets were acquired with *3dMDface* photogrammetric systems.[1]

2.3.2 Results cross-validation

We analyzed average errors made by the automatic landmarking algorithm by using a cross-validation procedure. Each face from the training set was excluded iteratively and the remaining training set was used to landmark the left out face. We used the Euclidean distance between the manually placed landmarks and the automatically found positions to measure landmarking error. We calculated average landmarking error per landmark and results are displayed in Table 2.1. Figure 2.3 shows automatic landmarking positions for all images in the sample.

It is apparent from Table 2.1 and Figure 2.3 that there is considerable variation between landmarks. Landmarks that perform best lie in the eyes (landmarks 1-4 and 8-10) and nose (landmarks 7 and 12-16) and the left and right corners of the mouth (landmarks 20 and 18). Landmarks that are structurally localized poorly and have many outliers are the center of the nose bridge (landmark 5), the forehead (landmark 6), lower lip bottom (landmark 19) and the chin dimple edge (landmark 21). Finally, the upper lip top center (landmark 17) varies greatly between the two data sets.

Inaccurately positioned landmarks include landmarks 6 (eye brows upper limit), 5 (brow ridge center), 19 (lower lip bottom center) and 21 (mouth right corner) (see again figure 2.3). For each of these landmarks the surrounding area showed little contrast in either texture or local shape, and as such, the layers used in our algorithm have difficulty in providing information about those landmarks. While landmark 5 is clearly placed on the correct vertical line, taking advantage of the topography of the ridge, its vertical placement varies strongly. Landmarks 6 and 21 show neither vertical nor horizontal edges and show great landmarking variability. Landmark 19 also has less clear boundaries, especially for the texture layer in which the edge of the lip is often unclear.

2.3.3 Importance of features

To assess the impact on performance of individual feature layers, we executed the algorithm using each of the feature layers separately. Results are shown in Table 2.2. When inspecting the results per feature layer, there is no clear-cut pattern in the performance of feature layers across studies, although there are some distinct differences in each data set within studies. In the *MeIn3D* dataset, the photographic and heightmap layers both perform worse than the trio of derivatives and Laplacian of Gaussian layers. In the *TwinsUK* data set, however, we see that the texture feature has comparable performance to the other features.

Contribution of each feature layer by comparing the "with feature" and "without feature" setup

Table 2.1: Mean distance and (standard deviation) in mm per landmark to the training data for the two data sets. Distances <2 mm are shown in bold.

	Description	[TwinsUK]		[MeIn3D]		Mean	STASM
1	left eye outer corner	2.4	(3.7)	2.0	(2.3)	2.2	2.2
2	left eye top	**1.7**	(1.0)	**1.2**	(0.6)	**1.4**	1.5
3	left eye inner corner	**1.8**	(1.3)	**1.2**	(0.6)	**1.5**	2.7
4	left eye bottom	2.3	(1.4)	**1.4**	(2.0)	**1.8**	1.5
5	brow ridge center	2.4	(1.6)	3.0	(2.2)	2.7	
6	eye brows upper limit	6.5	(4.8)	4.8	(3.8)	5.7	
7	nose tip	2.1	(1.2)	2.4	(1.4)	2.2	1.9
8	right eye inner corner	**1.5**	(0.8)	**1.4**	(1.6)	**1.4**	2.2
9	right eye top	**1.8**	(1.4)	**1.6**	(1.0)	**1.7**	1.8
10	right eye outer corner	2.2	(1.5)	**1.7**	(0.9)	2.0	2.6
11	right eye bottom	**1.7**	(1.3)	**1.5**	(2.4)	**1.6**	2.2
12	nose right limit	**1.1**	(0.6)	**1.2**	(0.7)	**1.1**	3.5
13	nose lower right	**1.4**	(0.7)	**1.4**	(1.0)	**1.4**	1.7
14	nose bottom	**1.2**	(0.7)	**1.2**	(0.5)	**1.2**	
15	nose lower left	**1.5**	(0.8)	**1.7**	(0.8)	**1.6**	1.6
16	nose left limit	**1.2**	(0.7)	**1.1**	(0.6)	**1.2**	3.6
17	upper lip top center	**1.2**	(0.8)	2.7	(3.2)	2.0	1.9
18	mouth right corner	2.0	(1.7)	**1.5**	(0.9)	**1.7**	2.2
19	lower lip bottom center	2.5	(1.6)	4.1	(4.5)	3.3	2.3
20	mouth left corner	**1.9**	(1.2)	2.2	(4.7)	2.0	2.2
21	chin dimple edge	3.3	(3.4)	8.5	(6.8)	5.9	3.2
Mean 21 landmarks		**2.1**		**2.3**		**2.2**	
Mean 18 Stasm landmarks		**1.9**					**2.3**

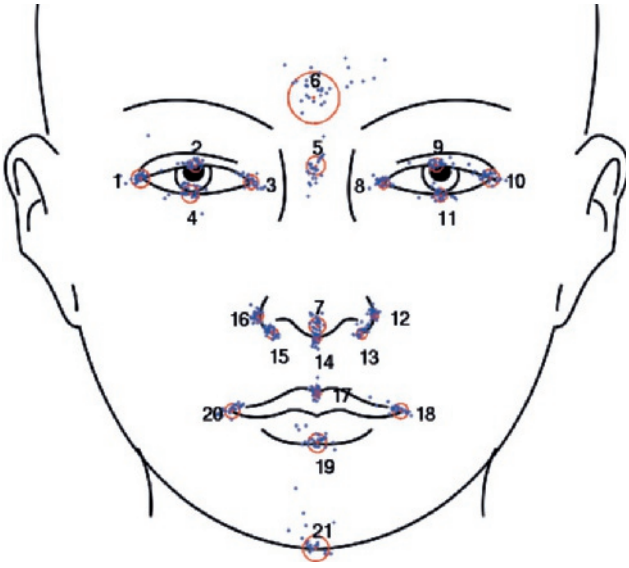
is visualized in a heatmap shown in Figure 2.4. In both datasets every feature is essential for some landmarks. For landmark *lower lip bottom center* height map is nonessential in the *MeIn3D* data and for *chin dimple edge* height map is nonessential in *TwinsUK*. In general, importance patterns agree across data sets. However, for some landmarks importance differs such as landmarks 1, 11, 15 and 19.

Table 2.2: Mean distance in mm per landmark per feature layer for the *MeIn3D* (left column) and *TwinsUK* (right column) data set. Distances <2 mm are shown in bold.

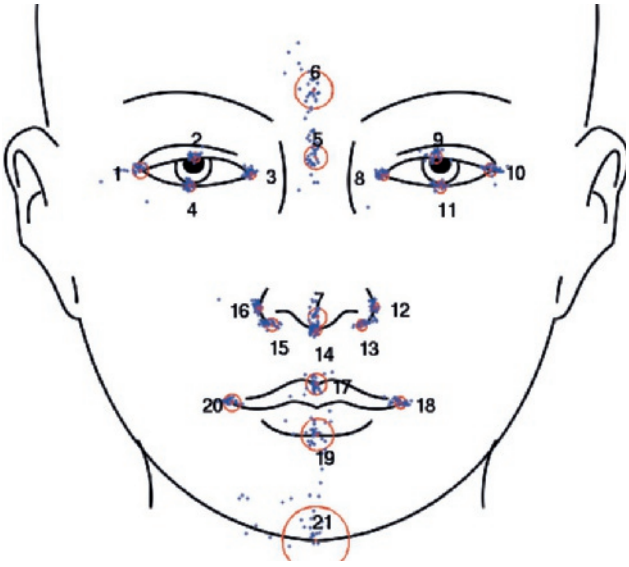
Landmark	Photo		Heightmap		Der x		Der y		LoG	
1	2.1	**1.6**	**1.5**	**1.5**	**1.6**	**1.4**	**1.7**	**1.4**	**1.8**	**1.6**
2	**1.1**	**1.9**	**1.1**	**1.6**	**1.4**	**1.6**	**1.3**	**1.7**	**1.5**	**1.5**
3	**1.4**	**2.0**	**1.6**	**1.8**	**1.0**	2.2	**1.1**	**2.0**	**1.2**	**1.8**
4	**1.2**	**1.2**	**1.2**	**1.2**	**1.2**	**1.6**	**1.5**	**1.6**	**1.5**	2.1
5	3.1	2.1	2.4	**1.7**	2.2	2.2	2.3	2.3	2.2	2.4
6	5.1	4.1	4.9	4.2	4.5	3.5	4.3	3.6	4.3	4.8
7	2.6	**1.9**	2.4	2.3	2.1	2.2	2.5	**1.9**	2.5	2.2
8	**1.5**	**1.8**	**1.7**	**1.8**	**1.1**	**1.7**	**1.3**	**1.5**	**1.3**	**1.5**
9	**1.3**	**1.8**	**1.7**	**1.4**	**1.5**	**2.0**	**2.0**	**1.7**	**2.0**	**1.7**
10	**1.6**	**1.3**	**1.6**	**1.3**	**1.8**	**1.8**	**1.6**	**1.7**	**1.6**	**2.0**
11	**1.6**	**1.6**	**1.5**	**1.3**	**1.4**	**1.9**	**1.3**	**1.5**	**1.3**	**1.5**
12	**1.3**	**1.3**	**1.2**	**1.3**	**1.1**	**0.9**	**1.2**	**0.9**	**1.2**	**1.0**
13	**1.4**	**1.2**	**1.6**	**1.2**	**1.3**	**1.3**	**1.4**	**1.4**	**1.4**	**1.4**
14	**1.2**	**1.2**	**1.2**	**1.4**	**1.3**	**1.4**	**1.3**	**1.2**	**1.2**	**1.1**
15	**1.3**	**1.0**	**1.2**	**1.1**	**1.5**	**1.3**	**1.7**	**1.3**	**1.6**	**1.4**
16	**1.0**	**1.3**	**1.0**	**1.4**	**1.1**	**1.1**	**1.0**	**1.2**	**1.2**	**1.3**
17	2.9	**1.5**	2.3	**1.7**	**1.6**	**1.3**	**1.8**	**1.1**	**1.7**	**1.3**
18	2.0	**1.5**	2.1	**2.0**	**1.4**	**2.0**	**1.4**	**1.9**	**1.4**	2.1
19	6.3	**1.6**	6.9	2.1	3.8	2.1	3.4	2.0	3.2	2.3
20	2.5	**1.2**	2.4	2.4	**1.4**	**1.8**	**1.8**	**1.9**	**1.8**	**1.9**
21	9.0	3.6	9.3	5.4	8.9	3.5	7.8	3.7	7.6	5.1
Mean	2.5	**1.8**	2.4	**1.9**	2.1	**1.8**	2.1	**1.8**	2.1	**2.0**

2.3.4 Results active shape model

Results of the ASM method comparison are given in Table 2.1 in the STASM column. Landmarks 5, 6 and 14 did not have a correspondence in the landmark set of the STASM algorithm and are not reported. Two subjects were not included in computing mean distance as the STASM method was unable to localize the face properly. Mean distances for the remaining samples are given. For three landmarks in the left eye (1: 0.2mm, 2: 0.2 mm, 4: 0.8mm), *lower lip bottom center*: 0.2mm, and *chin*



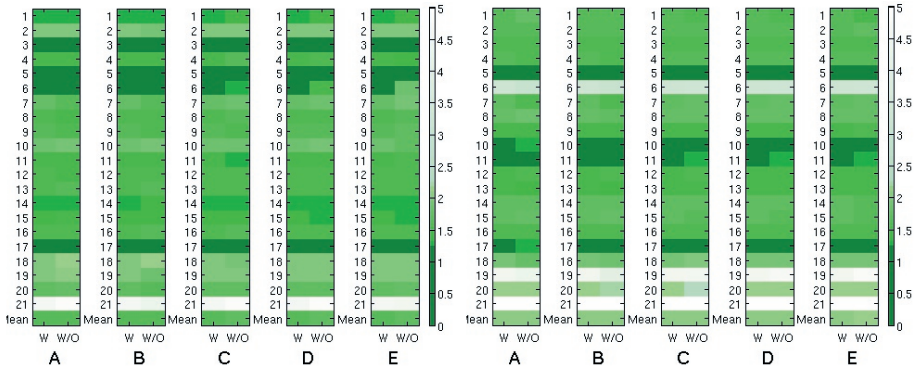
(a) TwinsUK



(b) MeIn3D

Figure 2.3: Schematic overview of (to scale) automatic landmarking results of the *TwinsUK* and *MeIn3D* data sets. Mean distance to training data is represented as a red circle.

dimple edge, the ASM performed better. The algorithms showed the same accuracy for one landmark (*right eye top*). Our algorithm outperforms STASM algorithm by 0.1-2.4mm for the remaining 12 landmarks. The mean error of STASM was 2.3 mm compared to 1.9 mm of our method.



(a) TwinsUK (b) MeIn3D

Figure 2.4: Contribution overview for each feature layer in the *TwinsUK* and *MeIn3D* data sets, with left column 'w' with this layer, right column 'w/o' without this layer. A: photographic, B: heightmap, C: derivative with respect to X-axis, D: derivative with respect to Y-axis, E: heightmap Laplacian of Gaussian. Distances are in mm.

2.3.5 Results heritability

Heritability as a measure of accuracy has the advantage that it is objective in the sense that rater error can be assessed from a single dataset by comparing expected similarity by degree of relationship with observed similarity. For this evaluation we manually labeled the full *twinUK* data set to compare human with automatic performance. We computed heritability of all 210 pairwise distances and results for 20 selected distances are shown in table 2.3, where they are compared to estimates based on manually placed landmarks. Results for all distances are reported in the appendix.

Table 2.3 shows results sorted by heritability of the automatic procedure. For these, heritability of manually labeled faces show very similar heritabilities. The magnitude of these heritabilities agrees with previously reported anthropometric measurements.[9] For some distances, manual labeling outperformed automatic landmarking substantially, for example *d_1_19* for which manual labeling resulted in heritability of 68% and the automatic procedure achieved 32% (see appendix).

2.4 Discussion

We present an automated approach for the landmarking of human facial 3D surface data. Our results indicate that for some landmarks excellent accuracy is achieved whereas certain landmarks showed lower accuracy. Accuracy was measured in two ways. Absolute discrepancy between automatically and manually placed landmarks are below 2 mm for most landmarks. Heritability estimates show that for the highest heritabilities according to the automatic procedure it could not be outperformed by a human rater (Table 2.1). For certain distances manual performance in terms of heritability was much better than automatic performance.

The heritability results imply that arguably all useful and accurate features were learned and used

Table 2.3: 20 most heritable distances in the TwinUK data set. *age* effect of age, σ_1 residual error, σ_2 random effects component. h^2 heritability estimate. Columns indexed *gt* contain estimates for manually placed landmarks. Other columns describe the automatic procedure. d_i_j denotes euclidean distance between node i and j.

Distance	**age**	σ_1	σ_2	h^2	**age**$_{gt}$	σ_{1gt}	σ_{2gt}	h^2_{gt}
d_7_12	0.00	1.24	1.72	0.66	-0.01	1.42	1.54	0.54
d_13_18	0.04	2.32	2.91	0.61	0.04	2.67	2.66	0.50
d_12_15	0.04	1.19	1.47	0.60	0.03	1.32	1.57	0.58
d_3_18	0.04	2.38	2.91	0.60	0.04	2.59	2.68	0.52
d_7_18	0.04	2.34	2.73	0.58	0.04	2.58	2.59	0.50
d_7_15	0.05	1.26	1.44	0.57	0.04	1.27	1.49	0.58
d_17_18	0.08	1.81	2.06	0.56	0.09	1.84	1.87	0.51
d_16_18	0.06	1.57	1.78	0.56	0.05	2.29	0.00	0.00
d_15_18	0.06	1.54	1.73	0.56	0.05	1.53	1.59	0.52
d_15_16	0.07	1.78	1.99	0.56	0.06	1.72	2.02	0.58
d_12_16	0.04	1.48	1.65	0.55	0.03	1.45	1.72	0.59
d_14_15	0.05	1.48	1.64	0.55	0.04	1.52	1.69	0.55
d_12_18	0.03	2.16	2.38	0.55	0.03	2.31	2.38	0.51
d_12_13	0.00	1.18	1.28	0.54	-0.01	1.18	1.32	0.56
d_14_18	0.04	1.89	2.04	0.54	0.04	2.04	1.99	0.49
d_12_14	0.02	1.34	1.43	0.53	0.01	1.47	1.38	0.47
d_14_16	0.06	1.70	1.81	0.53	0.04	1.62	1.82	0.56
d_19_21	-0.02	2.72	2.89	0.53	0.00	2.80	2.63	0.47
d_13_15	0.05	1.31	1.38	0.53	0.04	1.26	1.44	0.56
d_7_16	0.05	1.54	1.60	0.52	0.04	1.43	1.61	0.56
d_1_19	-0.01	2.47	1.71	0.32	-0.07	1.14	1.67	0.68

to place certain landmarks whereas for some landmarks such pertinent features were not available in the data layers derived in the algorithm. One advantage of our algorithm is that we can easily add new features that can be used to improve landmarking accuracy for thus far lacking landmarks. The combination of new feature and Gabor transform thereby implicitly defines a new filter for which the nature of the transformation implies a clear interpretation. We believe that adding well chosen layers can improve accuracy. For example, curvature and transformations thereof could potentially increase

accuracy for landmarks that have performed badly thus far. In summary, the distances of landmark results to true positions in the training data generally lie between 1-2 mm, which is in line with other methods.[13]

When we compare feature layer performance, it turns out that importance of feature layers is data set specific. A second finding is that features are rarely non-essential, *i.e.* their omission rarely improves landmarking performance. In our implementation this is an expected finding as we measure template similarity by the correlation of wavelet responses across all features thereby averaging out non-essential information. Both findings underline robustness of our algorithm and the fact that robust landmarking strategies need adaptive elements. Heuristic procedures tailored to specific data sets are expected to potentially perform much worse on new data sets.[13] We have not yet fully explored the aspect of feature layer selection which is a plan for future research. This includes selection and weighing of layers.

We have chosen the EBGM for automatic landmarking as it is well studied and shows good performance with small training samples. In principle, other methods such as active shape models (ASMs)[18] could be used and we acknowledge that ASMs exploit information about variability in the population better than EBGM. This is a possible explanation for the results found for the left eye where the ASM outperforms our method. Here, the ASM can make use of knowledge about joint placement of landmarks whereas our implementation locates landmarks almost independently. ASMs explore the search space in the direction of principal components (PCs) of the landmark space whereas EBGM searches around the population mean.[16] We believe that PCs can contribute to accurate landmarking by efficiently restricting the search space and we plan to investigate strategies to incorporate such prior information into our algorithm.

Apart from the issue of lack of usable features, the quality of landmark detection is limited by the detail in both the resolution of the photographs associated with the model and the amount of vertices in the model itself. Due to low 3D detail, highly informative 3D features such as eye corners and the edges of the eyelids and mouth are often not prominent or consistent in our data sets. Ideally, areas such as the corners of the eye should be clearly distinguishable in both 2D and 3D, this is however often not the case due to blurry textures and blocky 3D features. Some of our results are strongly influenced by the 3D triangulation technique used. First, as the 3D model is constructed from multiple photographs taken from different positions, there may be texture issues that become especially apparent in concave areas such as the eye corners. In those areas, multiple photographic source images meet during the triangulation process and are merged. This may result in sudden gradients or jagged tears. Second, triangulation of transparent areas such as the eye lens or semi-transparent features such as eyelashes remains difficult. In all these cases unpredictable artifacts in both 2D and 3D may arise. Differences in the quality of raw images make it difficult to compare accuracy between papers. Raw image quality should also be used to understand landmarking performance. For example, the issues mentioned above explain why eye corners are less accurately labeled than, say, nose features in our data sets.

Another issue is the possibility of occlusion of certain landmarks, *e.g.* the base of the nose between the nostrils, which may be occluded by the nose tip in certain (aquiline) nose shapes. Although our algorithm performed robustly for the data sets we used, potentially occluded landmarks can be a problem. By adding a projection that reliably reveals a thus far occluded landmark our algorithm can be extended to such use cases with very little modifications.

In comparison with a state of the art implementation of an ASM (STASM) our algorithm has shown superior results for the goal of landmarking with regard to ease of use and accuracy. Our method was able to perform better or comparable to ASM with a training set of 30 samples in contrast to 3000+ samples that were used for the ASM implementation. In another example, ASMs required up to 240 manually annotated training samples for 40 subjects.[18] This represents a considerable practical advantage of our approach as it greatly reduces the time that is needed to create a completely new set of landmarks or to make ad-hoc corrections to already landmarked data sets, which would typically take less than half an hour to complete. Our comparison result cannot be fully generalized as the ASM model was not optimized for the 3D surface case but the training complexity would remain a distinct advantage.

The non-heuristic nature of our approach in comparison to certain existing methods,[13] potentially allows for applications of our algorithm to other 3D surface data apart from faces. A practically important advantage of the map projection is that it is possible to manually place all 3D landmarks in a single image. The inclusion of additional landmarks, such as ears, can also be easily achieved.

Conclusions

The proposed method for automatic landmarking of facial 3D surface data requires little investment in the training phase (ca. 30 training samples of 1 minute each for 21 landmarks) to automatically landmark 3D faces in a single iteration. Our experiments show good performance (1-2 mm distance for most landmarks) over faces of difference quality, gender and ethnic background and the algorithm can be easily and quickly re-trained when a different set of landmarks is required. These properties are important for the landmarking of large medical 3D (facial) data sets.

References

[1] *3dMD — 3D Imaging Systems and Software*. URL: `http://www.3dmd.com/` (visited on 06/12/2014).

[2] Brunilda Balliu et al. "Classification and Visualization Based on Derived Image Features: Application to Genetic Syndromes". In: *PloS one* 9.11 (2014), e109033. URL: `http://dx.plos.org/10.1371/journal.pone.0109033` (visited on 05/28/2015).

[3] Stefan Boehringer et al. "Automated syndrome detection in a set of clinical facial photographs". In: *American Journal of Medical Genetics Part A* 155.9 (2011), pp. 2161–2169. ISSN: 1552-4833.

[4] Stefan Boehringer et al. "Genetic determination of human facial morphology: links between cleft-lips and normal variation". In: *European Journal of Human Genetics* 19.11 (2011), pp. 1192–1197. ISSN: 1018-4813.

[5] Stefan Boehringer et al. "Syndrome identification based on 2D analysis software". In: *European Journal of Human Genetics: EJHG* 14.10 (Oct. 2006), pp. 1082–9. ISSN: 1018-4813. DOI: 5201673. URL: `http://www.ncbi.nlm.nih.gov/pubmed/16773127` (visited on 11/13/2008).

[6] Timothy F. Cootes et al. "Active shape models-their training and application". In: *Computer vision and image understanding* 61.1 (1995), pp. 38–59. URL: http://www.sciencedirect.com/science/article/pii/S1077314285710041 (visited on 06/05/2014).

[7] Peter Hammond et al. "Discriminating Power of Localized Three-Dimensional Facial Morphology". In: *The American Journal of Human Genetics* 77.6 (Dec. 2005), pp. 999–1010. ISSN: 0002-9297. DOI: 10.1086/498396. URL: http://www.sciencedirect.com/science/article/pii/S0002929707633849 (visited on 02/05/2014).

[8] Peter Hammond et al. "Fine-grained facial phenotype–genotype analysis in Wolf–Hirschhorn syndrome". en. In: *European Journal of Human Genetics* 20.1 (Jan. 2012), pp. 33–40. ISSN: 1018-4813. DOI: 10.1038/ejhg.2011.135. URL: http://www.nature.com/ejhg/journal/v20/n1/full/ejhg2011135a.html (visited on 02/05/2014).

[9] L. A. P. Kohn. "The Role of Genetics in Craniofacial Morphology and Growth". In: *Annual Review of Anthropology* 20 (Jan. 1991), pp. 261–278. ISSN: 0084-6570. URL: http://www.jstor.org/stable/2155802 (visited on 07/16/2014).

[10] Fan Liu et al. "A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans". In: *PLoS Genetics* 8.9 (2012), e1002932. ISSN: 1553-7404.

[11] Stephen Milborrow and Fred Nicolls. "Active shape models with SIFT descriptors and MARS". In: *VISAPP* 1.2 (2014), p. 5. URL: http://www.dip.ee.uct.ac.za/~nicolls/publish/sm14-visapp.pdf (visited on 06/05/2015).

[12] Lavinia Paternoster et al. "Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in {\emph PAX3} Associated with Nasion Position". In: *The American Journal of Human Genetics* 90.3 (2012), pp. 478–485. URL: http://www.sciencedirect.com/science/article/pii/S000292971200002X (visited on 06/05/2014).

[13] Shouneng Peng et al. "Detecting Genetic Association of Common Human Facial Morphological Variation Using High Density 3D Image Registration". In: *PLoS computational biology* 9.12 (2013), e1003375. URL: http://dx.plos.org/10.1371/journal.pcbi.1003375.g004 (visited on 06/05/2014).

[14] Harald J. Schneider et al. "A Novel Approach to the Detection of Acromegaly: Accuracy of Diagnosis by Automatic Face Classification". In: *J Clin Endocrinol Metab* (Apr. 2011), jc.2011–0237. DOI: 10.1210/jc.2011-0237 . URL: http://jcem.endojournals.org/cgi/content/abstract/jc.2011-0237v1 (visited on 05/06/2011).

[15] Tobias Vollmar et al. "Impact of geometry and viewing angle on classification accuracy of 2D based analysis of dysmorphic faces". In: *European Journal of Medical Genetics* 51.1 (2008), pp. 44–53. ISSN: 1769-7212. DOI: S1769-7212(07)00104-8. URL: http://www.ncbi.nlm.nih.gov/pubmed/18054308 (visited on 11/13/2008).

[16] Laurenz Wiskott et al. "Face recognition by elastic bunch graph matching". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19.7 (1997), pp. 775–779. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=598235 (visited on 06/05/2014).

[17] Shizhong Xu. *Principles of Statistical Genomics*. en. Springer, Sept. 2012. ISBN: 9780387708072.

[18] Dianle Zhou, Dijana Petrovska-Delacretaz, and Bernadette Dorizzi. "Automatic Landmark Location with a Combined Active Shape Model". In: *Proceedings of the 3rd IEEE International Conference on Biometrics: Theory, Applications and Systems*. BTAS'09. Piscataway, NJ, USA: IEEE Press, 2009, 49–55. ISBN: 978-1-4244-5019-0. URL: `http://dl.acm.org/citation.cfm?id=1736406.1736414` (visited on 06/19/2014).

3

Ensemble landmarking of 3D facial surface scans

Markus A. de Jong
Pirro Hysi
Tim Spector
Wiro Niessen
Maarten J. Koudstaal
Eppo B. Wolvius
Manfred Kayser
Stefan Böhringer

Abstract

Landmarking of 3D facial surface scans is an important analysis step in medical and biological applications, such as genome-wide association studies (GWAS). Manual landmarking is often employed with considerable cost and rater dependent variability. Landmarking automatically with minimal training is therefore desirable.

We apply statistical ensemble methods to improve automated landmarking of 3D facial surface scans. Base landmarking algorithms using features derived from 3D surface scans are combined using either bagging or stacking. A focus is on low training complexity of maximal 40 training samples with template based landmarking algorithms that have proved successful in such applications. Additionally, we use correlations between landmark coordinates by introducing a search strategy guided by principal components (PCs) of training landmarks.

We found that bagging has no useful impact, while stacking strongly improves accuracy to an average error of 1.7mm across all 21 landmarks in this study, a 22% improvement as compared to a previous, comparable algorithm. Heritability estimates in twin pairs also show improvements when using facial distances from landmarks.

Ensemble methods allow improvement of automatic, accurate landmarking of 3D facial images with minimal training which is advantageous in large cohort studies for GWAS and when landmarking needs change or data quality varies.

3.1 Introduction

Interest in facial analysis has recently surged in genetic and genome-wide association studies (GWASs), partly due to the availability of large cohorts and partly due to availability of efficient surface scanning.[18] The aim of such studies is to explain phenotypic variation as a first step in understanding the genetic basis of the human face.[4,7,8,10,18,20,22,23] This situation contrasts with facial analysis in clinical genetics in which samples sizes are usually much smaller. In the clinical application, shape differences between groups tend to be large[5,14,21,24] within small cohorts, whereas in population based applications such as GWASs shape differences due to genetic variation are usually small.[4,18,20] As landmarking structure and input data vary across studies,[3,14,18] and as such require manual retraining of landmarking algorithms, both applications benefit from low training complexity.

Promising results for landmarking algorithm accuracy have been demonstrated by several landmarking approaches so far,[16,19,27] some heavily depending on heuristics.[13,17] Still, it is unclear whether strengths of individual algorithms are complementary, *i.e.* whether they can be combined to generate yet more accurate landmarking data. In previous work, we showed that different data transformations make additional information available to standard wavelet-based methods.[16] However, we noted two drawbacks that we overcome with the present study. Firstly, our previous approach performs unsatisfactorily for landmarks in areas with little structural information such as the cheeks or the chin region. Secondly, the choice of transformations that we used as input for our algorithm was not systematic or weighted, leaving open the question of optimality.

To address the first problem, we note that the distribution of landmark positions in the (training) population may provide additional information about the landmarks with little structural information. Such information can be exploited by using principal component analysis (PCA) of the landmark space such as used by active shape models.[9] The second problem poses a model selection problem wherein information from different input data transformations, or features, needs to be weighted and selected for each individual landmark.

In this study, we explore model selection of input features in combination with information from PCA of population coordinates under the constraint of small training samples. We employ the statistical ensemble methods of bagging and stacking to integrate all landmarking information into a single landmarking method. Model selection is performed as an intrinsic feature of the stacking combination technique.[11]

In a broad sense, ensemble methods have been employed in landmarking algorithms before. Elastic bunch graph matching (EBGM),[27] the method used for most of the base landmarking methods in this and our previous paper, can be viewed as an ensemble method as it integrates a bank of varying wavelet filters into a single matching score per landmark. However, no weighting takes place.

More recently, deep learning technology has been used in the landmarking problem.[29] Deep learners can also be viewed as ensembles, where base learners are repeatedly integrated in each new layer of the network. However, deep learning methodology is not suitable for the smaller training sample sizes we consider here. For example, one study made use of 20,000 training samples.[29] We therefore do not consider deep learners and focus on combination methods suitable for small training samples.

The paper is organized as follows: first, we describe the new landmarking algorithm. In the next sections, we detail several landmarking experiments that are evaluated using either cross-validation or heritability and present the results. We conclude with a discussion.

3.2 Methods

The landmarking algorithm presented here combines a number of base landmarking algorithms into an ensemble. A base landmarking algorithm can be any algorithm that can propose a landmark position given new input data. To abbreviate, we will refer to an individual landmarking algorithm as a *landmarker* in the following. Averages or regression predictions are used to predict the final landmark from landmarks proposed by the base landmarkers. In our implementation, all base landmarkers are template based. A small number (typically 30 to 40) of training images is manually labeled by a rater from which base landmarkers extract templates in the training phase.

As a pre-processing step, 2D projections of the raw 3D surface data are derived. 3D information is retained in a heightmap that corresponds point-wise to a 2D texture, making the transformation one-to-one. A number of features are generated from this combined 2D data that serve as input for the base landmarking algorithms.

All base landmarkers are based on Gabor wavelet responses. Most algorithms target different features and work analogously to the EBGM algorithm with local search strategies. An algorithm with a global search strategy based on principal components (PCs) is added to the ensemble. The choice of base landmarkers is discussed later. The landmark search for the base algorithms is initialized at the population mean.

Projection and Data preprocessing

Projection of 3D surface data onto a 2D plane works by fitting an ellipsoid to the facial surface data and applying a Mercator map projection that results in a relief map.

The region of interest (ROI) of the frontal face is delimited by a standard sized square placed following the map projection (for an example of the ROI, see Figure 3.1, 1A). The size of the 2D features generated from the 3D surface is 200x200 pixels (40,000 pixels total).

Feature set

Three main features are created by using data components that correspond one-to-one per pixel: photographic or *texture* (Figure 3.1, 1A), heightmap (Figure 3.1, 2A) and curvature (Figure 3.1, 3A).

The main texture feature is created directly from the map projection using the original photographic information attached to the map projection.[16]

The main heightmap feature is based on the elevation levels with respect to the ellipsoid that were retained after the map projection.[16]

The third main component, curvature, is newly introduced and derived as follows: first, the curvature per 3D edge of the surface mesh is calculated by taking the mean normal of the first two principal

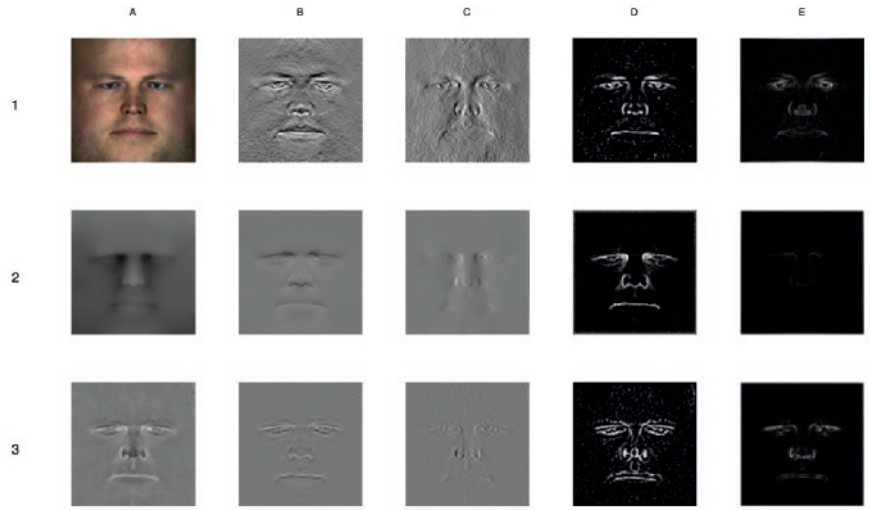


Figure 3.1: **Feature set overview.** Main feature displayed in the first column: (1A) texture, (2A) heightmap. (3A) curvature. The remaining columns show edge enhancements of the main features: (B) derivative over x-axis, (C) derivative over y-axis, (D) Laplacian of Gaussian filter, (E) Sobel filter. For illustration purposes, the face used in this image is that of author MadJ who was not a participant this study.

curvatures of the attached triangular surface patches. Secondly, within each patch of the triangulation, curvatures are computed by linear interpolation based on curvatures of the three related edges. Thirdly, curvatures are projected onto 2D using the projection derived above.

To enrich the number of available features, several data transformations that can be described as edge enhancements are applied to these components. These are: vertical and horizontal directional pixel derivatives, a Laplacian of Gaussian (LoG) filter and a newly introduced Sobel filter (Figure 3.1, columns B-E). In tests, each of these filters have shown good performances for non-overlapping subsets of landmarks. Any information overlap is expected to be removed through feature selection with ensemble methods. In total, 15 features are generated that form the input of the base landmarking algorithms.

Due to the one-to-one correspondence of pixels between features, training landmarks only have to be placed on a single feature image to be used for the complete set.

Base landmarking algorithms

Most base landmarking algorithms, or landmarkers, are based on the EBGM algorithm. In the training phase, a set of Gabor wavelets of different sizes and orientations is convoluted with all 15 individual features and the filter responses are extracted at the training landmarks, representing the templates. These responses are stored in a "bunch graph". In the landmarking phase, the set of Gabor wavelets is applied to a new image to be landmarked. Then, the bunch graph is read for a template search in

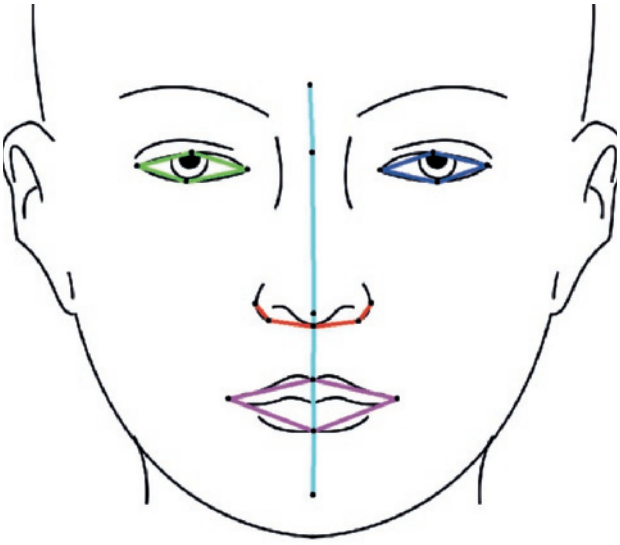


Figure 3.2: **Illustration of the 5 PC sub-groups.**

which responses from the training data are correlated with responses from the new image. The pixel coordinate for which maximum correlation is achieved, serves as the landmark prediction. Details of this procedure are given elsewhere.[16,26,27]

A first set of 15 base landmarkers is based on the individual features above, employing an EBGM algorithm on each. This set is augmented by two additional landmarkers. One base landmarker uses the sum of the wavelet responses of the 15 features for a template search. Another simply averages the final output coordinates of the 15 base landmarkers.

Principal Components

EBGM performs a local search around the starting position. To exploit correlations between coordinates of different landmarks, we introduce base landmarkers making use of PCA derived information.

For a given set of landmarks, PCA is performed on the landmark coordinates of training samples. The first two principal components (PCs) are used to direct a global search across these landmarks simultaneously.

Specifically, a neighborhood in the PC space is explored in a grid search on the first two PCs across all selected landmarks in the graph and across all features simultaneously, looking for a maximum combined correlation. The grid search is limited to a rectangular neighborhood, the size of which is defined by one standard deviation for that landmark in the training sample.

PCA exhibits high variability in loadings for small sample sizes.[15] For this reason, the facial graph is subdivided into five sub-graphs (Figure 3.2) to keep the number of variables small (4-6 landmarks) in relation with the sample size (30 or 40 training samples). The choice of sub-graphs is based on expected natural correlation between landmarks and symmetries. Otherwise, no systematic evaluation of possible sub-graphs was performed.

Whenever a sub-graph contains landmarks that overlap with a previously fitted sub-graph (i.e. the cyan group in Figure 3.2), the search is additionally penalized by the distance between overlapping landmarks.

Summary of base landmarkers

To summarize, we consider the following 18 base landmarkers:

- (1-15) The 15 landmarkers applying EBGM on individual features
- (16) The landmarker based on the sum of the wavelet responses from landmarkers 1-15
- (17) The mean of the final coordinates from landmarkers 1-15
- (18) The PC-based landmarker

Ensembles

Ensembles are uses to combine base landmarkers into a final landmarking algorithm. We consider two ensemble techniques: bagging, also known as bootstrap aggregating, and stacking, also known as stacked generalization.[11] Bagging has a smoothing property, whereas stacking has model selection properties by means of weighting base landmarkers.[11]

Bagging

The idea behind bagging is to create a large number of random sub-samples taken from a data set with replacement, called bags, after which fitting takes place in each of these bags. The final result is an average of the predictions of the individual models, with the intention that the average leads to a more stable predictor with less overfit than a single model fitted to the data would have.

In the present case, bags are created from the training data (30 facial scans) using 15 features. For each of these bags, base landmarkers are fitted that extract templates from the bags. Predicted landmark coordinates are averaged across the bags to give the final landmark position. Details of the bagging algorithm as used in this implementation are given in Algorithm 1.

Algorithm 1 30 item leave-one-out Bagging algorithm

procedure Bagging
  for each subject $s \in subjectset$ **do**
    for For n = 1:100 **do**
      $bootstrapsample_n$ = 29 random samples with replacement from $subjectset \backslash \{s\}$
      $EBGM_{bootstrapsample_n}$ = EBGM for all features trained with $bootstrapsample_n$
      $bootstrapresult_n = EBGM_{bootstrapsample_n}(s)$
    end for
    $finalresult_s = \frac{1}{100} \sum_{n=1}^{100} bootstrapresult_n$
  end for
end procedure

Stacking

In stacking, predictions from multiple low-level learning algorithms are used as input for a final combining top-level learning algorithm.[11] Stacking can be viewed as a feature selection procedure, as the final combiner typically weights the low-level algorithms. In our algorithm, the base landmarkers listed above are used for the low-level learning step. For the combination step, a least squares linear regression was applied. The three best predicting base landmarkers were selected according to the regression coefficients to create the final top-level predictor with these relative weights. Details of the stacking algorithm as used in this implementation are given in Algorithm 2.

Algorithm 2 40 item stacking algorithm

procedure Stacking
  for each $s \in subjects$ **do**
    $trainingsample = trainingdata \setminus \{trainingdata_s\}$
    $resultsample = resultdata \setminus \{resultdata_s\}$
    for each l in $landmarks$ **do**
      Perform linear regression: $trainingsample = \beta$*$resultsample + \epsilon$
      Predict l_s with β
    end for
  end for
end procedure

Heritability

Apart from cross-validation, we also used heritability to evaluate landmarking performance. Heritability is defined as the percentage of variation in a trait explained by genetic effects. Heritability can be estimated from families using a mixed effect model for which the variance of the random effect represents genetic effects and can be compared with residual variation.[28] We used twin data from the TwinsUK cohort for these analyses which included 37 monozygotic and 163 dizygotic twins. We estimated narrow sense heritability which assumes additive genetic effects for a number of features derived from landmark coordinates. To this end, we used a triangulation of the symmetrized mean graph to define a triangle structure. Then, coordinates were subjected to a Procrustes analysis using R package *shapes* and all distances between pairs of landmarks and all angles and areas of triangles were calculated for each of the samples. Heritabilities were calculated for each of these features as well as for landmark coordinates.

Heritabilities were visualized using importance plots which summarize heritabilities across features by computing a weighted average of the heritabilities for each point in the image. The weighting is linear in both the size of the individual heritability and the inverse distance of the center of the feature with the current point. Details of this procedure are given elsewhere.[2,12]

3.3 Experiments

The study and its experiments were conducted throughout 2016. All methods were performed in accordance with Erasmus MC guidelines and regulations according to which this study was not subject to evaluation by the medical ethical committee (http://www.ccmo.nl/en/non-wmo-research).

Data set

The data set used in the performance assessment of the presented algorithm is a random selection of 40 non-twin subjects from the *TwinsUK* cohort. The TwinsUK cohort consists of exclusive European descent.

The cohort consists of volunteers drawn from the general British population, unaware of any 3D studies scientific interests at the time of enrollment and gave fully informed consent under a protocol reviewed by the St. Thomas' Hospital Local Research Ethics Committee. Reference: PMID 23088889.

The *TwinsUK* dataset has models with ca. 1.5×10^5 points and textures with resolution of ca. 2,000x1,000 pixels. The data set was acquired with *3dMDface* photogrammetric systems.[1]

Data Availability

Due to privacy restrictions, raw data (i.e. facial 3D surface scans) cannot be made available for download. Subject to evaluation of a research proposal, the TwinsUK data set is made available by co-authors PH and TS.

Accuracy estimation

Cross-validations were performed to evaluate accuracy for a set of 21 landmarks. The different landmarkers were tested in leave-one-out experiments in which the ground truth consisted of a single manual labeling of the entire data set.

Additionally, we estimated heritability on the whole data set which can be done without knowing the ground truth.

Feature Set and Principal Components

All individual features and the PC-based predictions were all tested with a 40-item leave-on-out setup.

Bagging

Bagging was tested with a 30-item leave-one-out setup. Due to the large amount of iterations that were required (60,000), the experiment was performed on a computer cluster.

Stacking

Stacking of the base landmarkers was tested with a 40-item leave-one-out setup that included the complete feature set and PC predictions.

Results

Cross-validation

Results in this paper are compared to previous results, called the benchmark, as given by a previous version of our algorithm[16] that did not make use of ensemble learning. This earlier algorithm has been shown to outperform an active shape model based landmarking approach for most landmarks.[16]

Table 3.1 shows the results for each of the 15 base landmarkers obtained by EBGM from the respective features. Table 3.2 shows results for ensemble methods together with benchmark results from our previous algorithm. Both tables report Euclidean distance to the ground truth (training data) in mm. The final, stacked results are visualized in Figure 3.3.

	Texture					Heightmap					Curvature				
Landmark	*Ori*	*Dx*	*Dy*	*LoG*	*Sob*	*Ori*	*Dx*	*Dy*	*LoG*	*Sob*	*Ori*	*Dx*	*Dy*	*LoG*	*Sob*
1	*4.5*	*7.8*	*5.9*	*7.5*	3.7	*4.4*	*5.0*	*8.0*	*8.5*	*4.9*	*5.9*	4.0	3.2	3.6	*4.5*
2	3.6	*7.6*	3.9	*6.9*	*5.0*	3.9	*4.4*	*4.7*	*9.2*	3.2	3.9	3.5	2.6	3.7	3.3
3	3.8	*5.4*	3.0	3.5	3.2	3.0	2.6	*6.7*	*5.2*	3.5	3.0	2.4	2.4	3.3	2.6
4	3.7	*5.6*	*4.8*	*6.7*	3.7	2.8	2.0	*4.6*	*7.0*	3.6	*4.6*	*4.2*	1.9	2.5	3.6
5	*4.0*	2.3	*4.3*	*5.4*	*4.3*	*5.4*	3.6	*6.0*	*7.5*	*4.8*	*6.7*	3.8	*6.2*	*4.4*	*6.1*
6	*6.2*	*8.0*	*6.5*	*5.8*	*6.2*	*7.3*	*6.3*	*7.0*	*8.1*	*6.6*	*7.0*	*6.7*	*6.1*	*6.6*	*7.5*
7	2.3	*4.9*	3.0	3.1	2.6	2.2	2.0	*5.1*	*6.8*	2.6	2.0	2.3	1.9	2.0	2.3
8	3.2	*4.8*	*4.2*	*8.6*	3.5	3.2	3.0	*5.8*	*6.4*	*4.1*	2.9	2.9	*4.5*	2.5	2.6
9	*4.4*	*4.3*	3.5	*8.1*	*4.2*	*4.7*	4.0	*5.4*	*8.3*	3.8	*4.9*	3.6	3.1	3.6	3.4
10	3.9	*5.0*	*5.7*	*7.8*	*4.2*	2.9	3.0	*6.3*	*7.9*	*4.9*	*6.7*	2.6	2.2	2.4	2.2
11	3.2	4.0	3.8	*5.2*	3.9	2.4	2.2	3.6	*6.0*	3.2	2.8	3.0	2.4	2.9	2.2
12	1.9	*6.2*	2.6	2.0	2.3	3.0	2.8	*5.6*	2.1	2.0	2.1	1.9	2.9	2.9	2.6
13	2.3	2.5	*4.5*	2.7	2.1	2.7	2.0	*6.4*	2.3	2.1	2.0	2.1	2.2	2.3	2.3
14	2.2	2.3	2.4	3.9	2.4	3.3	3.2	2.8	3.9	3.3	2.3	2.2	2.3	2.7	2.3
15	2.0	2.5	3.6	2.3	2.1	1.8	1.8	3.9	1.9	1.8	2.0	2.1	1.9	1.8	1.9
16	1.9	2.6	2.9	2.3	2.1	2.1	2.4	*4.7*	2.3	2.1	2.0	2.3	2.2	2.3	2.0
17	3.9	*6.6*	3.3	*5.4*	2.5	*6.0*	*7.9*	*4.3*	*8.7*	*4.1*	3.1	2.5	3.0	3.5	*4.7*
18	*4.4*	*14.4*	*6.9*	*14.6*	*5.0*	3.8	3.4	*17.8*	*16.7*	*4.2*	*10.7*	*4.0*	1.8	1.8	3.1
19	*4.3*	*9.6*	*8.1*	*12.5*	3.7	*6.4*	*5.5*	*8.7*	*12.2*	3.6	*6.5*	*4.7*	*6.3*	*4.6*	*8.7*
20	*4.2*	*14.2*	*8.6*	*12.3*	*4.9*	3.0	*4.5*	*15.3*	*17.4*	*6.0*	*9.1*	3.0	*4.5*	2.6	*5.0*
21	*8.9*	*8.2*	*11.1*	*12.8*	*11.8*	*12.0*	*7.5*	*11.1*	*14.4*	*4.2*	*12.7*	3.3	*8.5*	*6.6*	11.3
mean	3.8	6.1	4.9	6.6	4.0	4.1	3.8	6.8	7.7	3.7	4.9	3.2	3.4	3.3	4.0
sd	1.6	3.5	2.3	3.8	2.1	2.3	1.8	3.7	4.4	1.3	3.1	1.1	1.9	1.4	2.5

Table 3.1: **Automatic landmarking results for 15 base landmarkers.** Results are reported in Euclidean distance to manual training data in mm, split by main feature (texture, heightmap, curvature) and sub-feature: Ori = no filter, Dx = derivative over x-axis, Dy = derivative over y-axis, LoG = Laplacian of Gaussian filter, Sob = Sobel filter. Distances <2mm are underlined, distances >4mm are in *italics*.

As concluded from our previous algorithm, Table 3.1 shows that the individual features are able to provide unique information for specific landmarks, implied by small distances for those landmarks.

Landmark	[Benchmark]		SoWR 15	Mean 15	PC	[Bagging]	[Stacking]	
1	2.4	(3.1)	2.4	3.1	*4.4*	3.7	1.8	(1.3)
2	1.7	(0.8)	2.4	2.8	2.8	2.4	2.1	(1.4)
3	1.8	(1.0)	1.9	1.9	2.6	1.8	1.6	(1.0)
4	2.3	(1.7)	1.7	2.1	2.7	1.4	1.5	(0.8)
5	2.4	(1.9)	2.9	2.4	2.8	2.2	2.0	(1.6)
6	6.5	(4.3)	*4.7*	3.3	*4.9*	*5.8*	3.0	(2.0)
7	2.1	(1.3)	1.7	1.9	*14.5*	3.0	1.4	(0.6)
8	1.5	(1.3)	2.0	2.2	3.2	1.5	1.9	(1.0)
9	1.8	(1.2)	2.7	2.7	*4.4*	2.8	2.3	(1.9)
10	2.2	(1.2)	1.7	2.4	3.9	3.0	1.6	(0.9)
11	1.7	(1.9)	1.6	1.6	3.1	1.6	1.3	(0.8)
12	1.1	(0.7)	1.9	1.7	*12.2*	1.6	1.3	(0.8)
13	1.4	(0.9)	2.0	1.6	*13.1*	1.5	1.5	(0.7)
14	1.2	(0.6)	2.2	1.7	*10.5*	2.2	1.5	(0.9)
15	1.5	(0.8)	1.6	1.5	*13.5*	1.5	1.3	(0.7)
16	1.2	(0.7)	1.7	1.9	*14.1*	2.0	1.7	(1.0)
17	1.2	(2.3)	1.8	2.9	*14.3*	2.6	1.5	(0.9)
18	2.0	(1.4)	1.4	*4.4*	*18.3*	2.8	1.4	(0.9)
19	2.5	(3.4)	3.1	3.4	*15.8*	3.6	2.1	(2.3)
20	1.9	(3.4)	1.8	*4.1*	*18.4*	3.0	1.8	(1.7)
21	3.3	(5.4)	3.1	*6.1*	2.1	*10.0*	2.0	(1.2)
mn	2.1		2.2	2.6	*8.6*	2.9	1.7	
sd	1.3		0.8	1.1	6.0	1.9	0.4	

Table 3.2: **Ensemble landmarking and PC results.** Results are reported in Euclidean distance to manual training data in mm. *Benchmark* represent results from the previous version of our algorithm.[16] Clarification of terms: *SoWR 15* = based on intermediate Summation of Wavelet Responses of 15 landmarkers. *Mean 15* = mean of final coordinates of 15 landmarkers. *PC* = results obtained by our principal component method. Distances <2mm are underlined, distances >4mm are in *italics*. Standard deviations are shown in parentheses.

When studying the results of the newly introduced main curvature feature, it can be seen this feature set improves results for many landmarks, especially 7 (nose tip) and 18 (left mouth corner).

The newly introduced Sobel filter sub-feature, shown in the same table in the leftmost column of each main feature, shows good results for landmark 13 (right nose corner) in the texture feature subset and landmark 12 (right nose outer edge) in the heightmap feature subset. Even though mean distance does not decrease much for those subsets, the Sobel filter contributes to coordinate stability by lowering standard deviations. Their symmetrical landmark partners (landmark 16 for 12 and landmark 15 for 13) also perform well but are still outperformed by other features.

The PC results are shown in Table 3.2 and are compared with our benchmark, results from our previous algorithm.[16] PC-based landmarker successfully improved results for difficult landmarks on the forehead (6) and chin (21), reducing distances from 6.5mm and 3.1mm to 3mm and 2.1mm respectively. Whilst PC application was especially focused on the forehead and chin, better results should also be attainable for the nose and mouth.

The bagging experiment did not improve results. This is most likely caused by the fact that

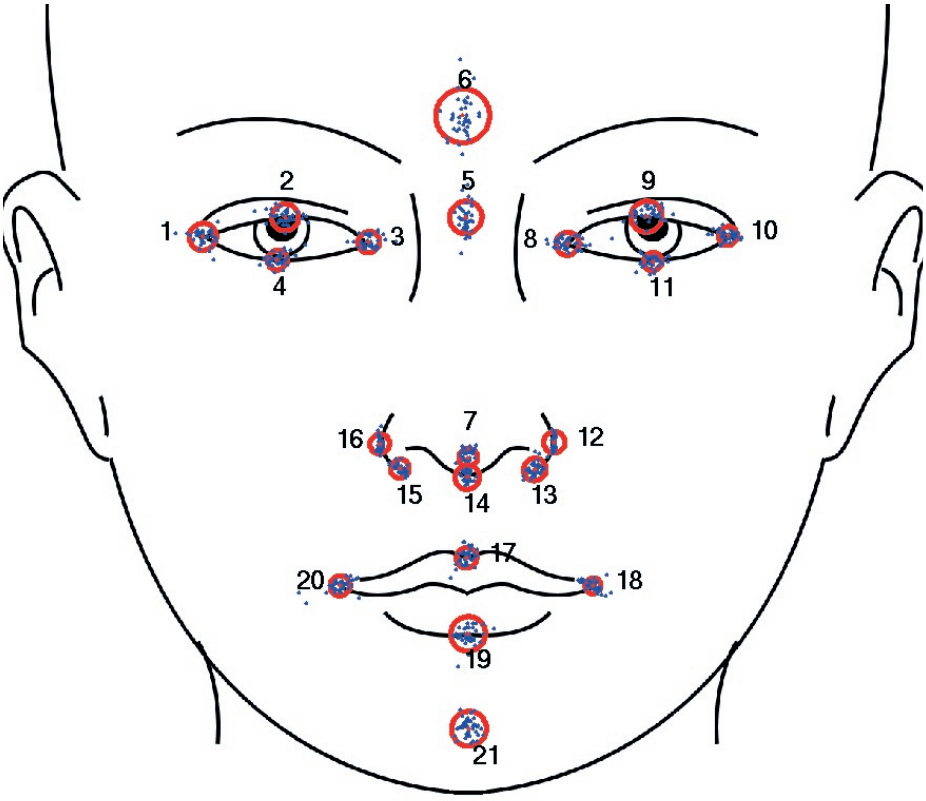


Figure 3.3: **Stacking final results**

• relative landmark result spread, all 40 final leave-one-out landmark results are plotted over each other.

O mean distance to the training landmarks.

pertubations to the data such as biological variation, the measure procedure, and labeling errors were roughly comparable across training samples and bagging could not smooth out any outliers due to atypical training samples. Besides these disappointing results, the large computational cost suggests limited use of bagging in landmarking.

Using the stacking algorithm, a significant mean improvement of 0.4mm across all 21 landmarks was achieved in comparison with our benchmark (2.1mm vs 1.7mm). A closer closer comparison shows better performance in all landmarks except 2, 8, 9, 12, 13, 14, 16 and 17. Overall, the stacking method algorithm is able to successfully optimize feature selection and is able to reduce distances. Furthermore, standard deviations are greatly reduced, leading to more reliable and stable landmarks.

Heritability

Heritability estimates for the extracted features are shown in Table 3.3 for the five most heritable features in each category. Heritabilities for all features are given as supplementary information. The highest observed heritability was 87% for the area of the triangle defined by landmarks 10, 12, and 18.

The best distance had heritability of 72%. Angles and coordinates had best heritabilities of 69% and 64%, respectively.

Feature	β_0	β_{age}	σ_1	σ_2	h^2
Coordinates					
c_12_x	-16.31	-0.01	0.76	1.01	0.64
c_1_x	-43.16	-0.00	1.05	1.31	0.61
c_18_x	-24.47	-0.01	1.36	1.44	0.53
c_3_x	-17.35	-0.02	0.93	0.95	0.51
c_13_x	-12.17	-0.02	0.80	0.74	0.46
Distances					
d_3_13	50.62	0.05	1.10	1.78	0.72
d_3_18	73.62	0.09	1.82	2.80	0.70
d_1_8	60.75	0.02	1.51	2.23	0.69
d_1_18	91.10	0.07	1.89	2.78	0.69
d_4_16	32.72	0.04	1.49	2.15	0.68
Areas					
ar_18_12_10	71.15	10.58	38.39	100.00	0.87
ar_8_7_12	527.87	0.60	40.85	50.10	0.60
ar_8_7_5	455.48	1.60	41.90	49.52	0.58
ar_14_13_7	122.52	0.35	14.82	13.30	0.45
ar_13_18_12	95.57	0.23	18.39	14.45	0.38
Angles					
an_18_12_10_b	2.02	0.00	0.06	0.09	0.69
an_18_12_10_a	0.69	-0.00	0.04	0.05	0.59
an_13_17_18_b	1.02	0.00	0.08	0.09	0.55
an_19_17_18_b	1.13	0.00	0.07	0.08	0.55
an_18_12_10_c	0.43	0.00	0.04	0.04	0.50

Table 3.3: **Heritabilities of geometric features.** β_0, β_{age} represent fixed effects of the model, σ_1, σ_2 are variances of the residual error and random effect, respectively.

Graphical summaries of heritabilities by means of importance plots are given in Figure 3.4. By comparing the overall summary 3.4 (S) with components C, D, R, and A it is apparent that distances contribute most to overall heritability. Heritabilities for all features except the raw coordinates are concentrated in the central area of the face. To analyze similarities within related individuals in the periphery, it is arguably better to work with the raw coordinates as indicated by Figure 3.4 (C).

When comparing these results with benchmark results, the best heritability for distances improved from 66% to 72%. In general, heritabilities improved by $\sim$ 5% when comparing the sorted lists for distances although the distances were not the same.

3.4 Discussion

In this paper, we evaluated ensemble methods to integrate information from several landmarkers (or base landmarking algorithms) in order to improve landmarking accuracy. This approach was motivated by experiences in previous landmarking efforts.[16]

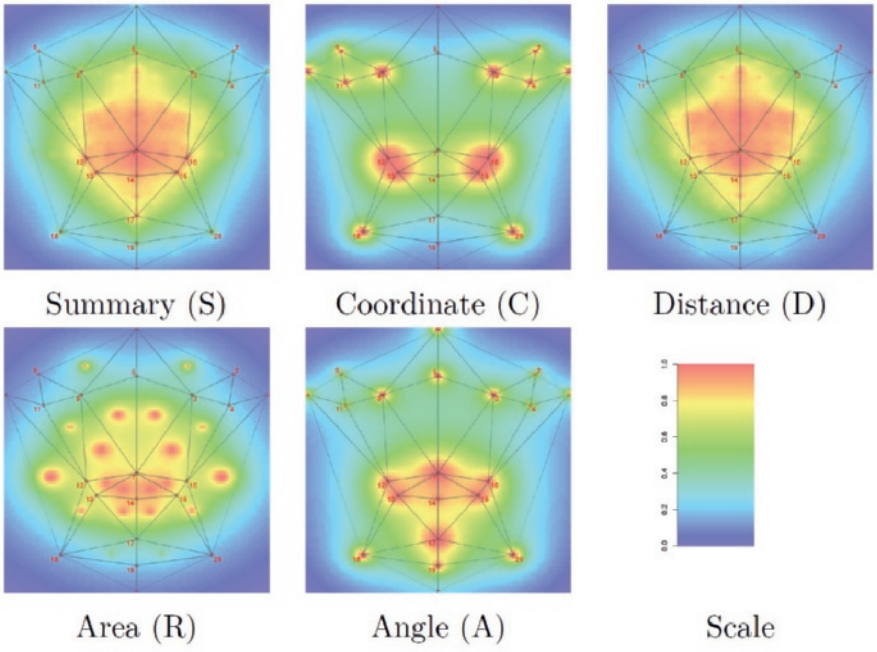


Figure 3.4: **Importance plots of heritabilities** of coordinates (C), distances (D), areas (R), angles (A), and a summary (S). Each color scale represents heritabilities which are re-scaled between 0 (blue) and the maximal heritability (red) for the respective feature.

By experimenting with different selections of features, it became apparent that features contribute only to a subset of landmarks. Additionally, some landmarks were poorly placed as revealed by inter-rater disagreement which was sometimes caused by atypical training samples. Ensemble methods can address both of these problems. Stacking can down-weigh features that are less relevant for a particular landmark and bagging can limit the influence of single training instances by smoothing predictions across bags.

Our results indicate that the composition of the training sample only has a small impact on labeling accuracy as bagging did not improve landmarking accuracy (Table 3.2). Moreover, this result justifies the use of small training samples as landmarking seems to be robust against changes in training example composition, as bagging contributes this type of variability into the landmarking algorithm. The stacking algorithm resulted in the overall best landmarking accuracy and performed best for almost all individual landmarks (Table 3.2). Any declines in accuracy for stacking in comparison with the benchmark can most likely be attributed to differences in methodology between both algorithms and 2D to 3D coordinate conversion. Nevertheless, the stacking experiment confirms that contributions of base landmarkers are indeed landmark specific and that a weighted combination can take advantage of this fact.

symmetrical landmarks agree within 0.4mm of accuracy for stacking, and usually within 0.2mm. Potential explanations for this symmetrical disagreement are asymmetries in the data, inaccuracies in preprocessing (ROI selection, projection), or random fluctuations due to non-deterministic steps in the

algorithm. These comparisons give a sense of the influence of these factors on labeling accuracy and they are roughly an order of magnitude smaller than the accuracies themselves.

In this work, we added new features to the previous algorithm: curvature as main feature and Sobel filter as sub-feature. All of these features did contribute to improve landmarking accuracy for subsets of landmarks. The base landmarker based on a PC guided search did improve landmarking accuracy for landmarks with little structural information by borrowing information from correlated landmarks. The stacking approach ensures that PC information is used for the appropriate landmarks. It therefore seems a sensible strategy to further enrich the number of available features to improve landmarking accuracy. On the other hand, the explicit need to define features is a disadvantage of our algorithm. Some features do not perform well for any landmark (e.g. Laplacian of Gaussian of the texture) and adding features that are too noisy will most likely decrease landmarking accuracy, despite stacking.

Deep learning offers an interesting alternative by working on raw data directly, thereby circumventing the need to specify features a-priori.[29] A disadvantage of deep learning approaches, however, is the need for big training samples. Up to a thousand-fold increase would be required in comparison to what we use in our current algorithm.[29] This contradicts with our aim to enable fast training of the landmarking algorithms, either for new data sets or for different sets of landmarks. A possible compromise could be to provide a limited number of features and add a network with a smaller number of layers than are used for deep networks trained on big sample sizes. We so far have focused on EBGM based base landmarkers as these can cope with small sample sizes. Using transformations learned from deep learning algorithms - an approach coined transfer learning[6] - could be a more flexible and generic than our current algorithm and would also retain the advantage of requiring small training samples. Such an approach could be more flexible and generic than our current algorithm and would also retain the advantage of requiring small training samples. It is our intention to investigate such possibilities in future research.

Potentially, large data sets might become available in the near future through consumer grade scanning devices and from social media resources. For such data, low training complexity might be less important. However, we believe that in research settings where data privacy is an important issue and data sets are often older, data specific methods with easy re-training will remain important in the future.

Heritability is an important aspect for genetic analyses. It is more likely to find genetic associations for highly heritable traits than for lesser heritable ones. Several of the estimated heritabilities range between 70% and 80%, values that are also seen in studies using manual landmarks,[25] although it is difficult to compare heritabilities across studies. We mainly use heritability as a benchmark that measures landmarking accuracy. Landmarking errors due to the algorithm contribute to residual variance of a measurement and thereby diminish heritability estimates. In general, estimated heritabilities improved in comparison with our previous iteration.[16] Distances were the most heritable traits in general and heritability was concentrated in the mid-face, which is a plausible finding. We believe that heritability is a valuable measure for landmarking accuracy when data is available that allows its estimation.

In this study, we present an improved landmarking algorithm for the human face that is based on ensembles and can incorporate an increasing number of features. Selection in the ensemble formation ensures that for a given landmark only useful information is gathered from base landmarkers which in

turn make use of specific features. This result is achieved with a low training complexity of 30 to 40 training samples. We were also able to tackle the problem of landmarks with little structural information by using a PC guided search. Overall we achieved an average accuracy of 1.7mm, a 22% improvement over our previous algorithm.

Furthermore, in comparison with another automated landmarking method with a comparable landmark set,[17] our algorithm showed better overall performance (2.6mm vs. 1.7mm for us). This positive comparison also holds when inspecting their best-performing individual landmarks: landmark 7 (tip of the nose) (1.6mm vs. 1.4mm for us), and landmark 13 (1.6mm vs. 1.5mm for us). Our results show that facial features can be extracted efficiently for large cohorts both in terms of time and cost and thereby enable research on facial morphology in such samples. This includes questions with respect to genetic mechanisms such as pursued in genome wide association studies (GWASs) and medical questions about normal variation, asymmetry, and classification.

References

[1] *3dMD — 3D Imaging Systems and Software*. url: `http://www.3dmd.com/` (visited on 06/12/2014).

[2] Brunilda Balliu et al. "Classification and Visualization Based on Derived Image Features: Application to Genetic Syndromes". In: *PloS one* 9.11 (2014), e109033. url: `http://dx.plos.org/10.1371/journal.pone.0109033` (visited on 05/28/2015).

[3] Stefan Boehringer et al. "Automated syndrome detection in a set of clinical facial photographs". In: *American Journal of Medical Genetics Part A* 155.9 (2011), pp. 2161–2169. issn: 1552-4833.

[4] Stefan Boehringer et al. "Genetic determination of human facial morphology: links between cleft-lips and normal variation". In: *European Journal of Human Genetics* 19.11 (2011), pp. 1192–1197. issn: 1018-4813.

[5] Stefan Boehringer et al. "Syndrome identification based on 2D analysis software". In: *European Journal of Human Genetics: EJHG* 14.10 (Oct. 2006), pp. 1082–9. issn: 1018-4813. doi: 5201673. url: `http://www.ncbi.nlm.nih.gov/pubmed/16773127` (visited on 11/13/2008).

[6] Philippe Burlina et al. "Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis". In: *Computers in Biology and Medicine* 82 (2017), pp. 80–86.

[7] Peter Claes et al. "Modeling 3D facial shape from DNA". In: *PLoS genetics* 10.3 (2014), e1004224.

[8] Joanne B Cole et al. "Genomewide association study of African children identifies association of SCHIP1 and PDE8A with facial size and shape". In: *PLoS genetics* 12.8 (2016), e1006174.

[9] Timothy F. Cootes et al. "Active shape models-their training and application". In: *Computer vision and image understanding* 61.1 (1995), pp. 38–59. url: `http://www.sciencedirect.com/science/article/pii/S1077314285710041` (visited on 06/05/2014).

[10] Jens Fagertun et al. "Predicting facial characteristics from complex polygenic variations". In: *Forensic Science International: Genetics* 19 (2015), pp. 263–268.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, 2001.

[12] Manuel Günther et al. "Reconstruction of images from Gabor graphs with applications in facial image processing". In: *International Journal of Wavelets, Multiresolution and Information Processing* 13.04 (2015), p. 1550019.

[13] Jianya Guo, Xi Mei, and Kun Tang. "Automatic landmark annotation and dense correspondence registration for 3D human facial images". In: *BMC bioinformatics* 14.1 (2013), p. 232.

[14] Peter Hammond et al. "Fine-grained facial phenotype–genotype analysis in Wolf–Hirschhorn syndrome". en. In: *European Journal of Human Genetics* 20.1 (Jan. 2012), pp. 33–40. issn: 1018-4813. doi: 10.1038/ejhg.2011.135. url: http://www.nature.com/ejhg/journal/v20/n1/full/ejhg2011135a.html (visited on 02/05/2014).

[15] I. T. Jolliffe. *Principal Component Analysis*. en. 2nd ed. Springer Science & Business Media, Mar. 2013. isbn: 978-1-4757-1904-8.

[16] Markus A de Jong et al. "An automatic 3D facial landmarking algorithm using 2D Gabor wavelets". In: *IEEE Transactions on Image Processing* 25.2 (2016), pp. 580–588.

[17] Shu Liang et al. "Improved detection of landmarks on 3d human face data". In: *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE. 2013, pp. 6482–6485.

[18] Fan Liu et al. "A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans". In: *PLoS Genetics* 8.9 (2012), e1002932. issn: 1553-7404.

[19] Stephen Milborrow and Fred Nicolls. "Active shape models with SIFT descriptors and MARS". In: *VISAPP* 1.2 (2014), p. 5. url: http://www.dip.ee.uct.ac.za/~nicolls/publish/sm14-visapp.pdf (visited on 06/05/2015).

[20] Lavinia Paternoster et al. "Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in {\emph PAX3} Associated with Nasion Position". In: *The American Journal of Human Genetics* 90.3 (2012), pp. 478–485. url: http://www.sciencedirect.com/science/article/pii/S000292971200002X (visited on 06/05/2014).

[21] Harald J. Schneider et al. "A Novel Approach to the Detection of Acromegaly: Accuracy of Diagnosis by Automatic Face Classification". In: *J Clin Endocrinol Metab* (Apr. 2011), jc.2011–0237. doi: 10.1210/jc.2011-0237 . url: http://jcem.endojournals.org/cgi/content/abstract/jc.2011-0237v1 (visited on 05/06/2011).

[22] John R Shaffer et al. "Genome-wide association study reveals multiple loci influencing normal human facial morphology". In: *PLoS genetics* 12.8 (2016), e1006149.

[23] Dimosthenis Tsagkrasoulis et al. "Heritability maps of human face morphology through large-scale automated three-dimensional phenotyping". In: *Scientific Reports* 7 (2017).

[24] Tobias Vollmar et al. "Impact of geometry and viewing angle on classification accuracy of 2D based analysis of dysmorphic faces". In: *European Journal of Medical Genetics* 51.1 (2008), pp. 44–53. issn: 1769-7212. doi: S1769-7212(07)00104-8. url: http://www.ncbi.nlm.nih.gov/pubmed/18054308 (visited on 11/13/2008).

[25] Seth M Weinberg et al. "Heritability of face shape in twins: a preliminary study using 3D stereophotogrammetry and geometric morphometrics". In: *Dentistry 3000* 1.1 (2013).

[26] Laurenz Wiskott and Christoph Von Der Malsburg. "Recognizing faces by dynamic link matching". In: *Neuroimage* 4.3 (1996), S14–S18. url: http://www.sciencedirect.com/science/article/pii/S1053811996900439 (visited on 06/05/2014).

[27] Laurenz Wiskott et al. "Face recognition by elastic bunch graph matching". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19.7 (1997), pp. 775–779. url: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=598235 (visited on 06/05/2014).

[28] Shizhong Xu. *Principles of Statistical Genomics*. en. Springer, Sept. 2012. isbn: 9780387708072.

[29] Zhanpeng Zhang et al. "Learning deep representation for face alignment with auxiliary attributes". In: *IEEE transactions on pattern analysis and machine intelligence* 38.5 (2016), pp. 918–930.

4

A clinical application in facial surgery - Three-dimensional orofacial soft tissue effects of mandibular midline distraction and surgically assisted rapid maxillary expansion: an automatic stereophotogrammetry landmarking analysis

Atilla Gül
Markus A. de Jong
Jan Pieter de Gijt
Eppo B. Wolvius
Manfred Kayser
Stefan Böhringer
Maarten J. Koudstaal

Abstract

Introduction: Research on mandibular midline distraction (MMD) is mostly performed using conventional methods. Concerning surgically assisted rapid maxillary expansion (SARME), more research is conducted with three-dimensional (3D) techniques. Research on bimaxillary expansion (BiMEx), combination of MMD and SARME, is reported sparsely. Main objective of this study is to provide a 3D evaluation of soft tissue effects following MMD and/or SARME.

Material and methods: From 2008 to 2013, non-syndromic patients who underwent MMD and/or SARME were included. Stereophotogrammetry records were taken at: pre-operative (T1), direct post-distraction (T2) and 1-year post-operative (T3). Analyses were performed with automatic 3D facial landmarking algorithm using 2D Gabor wavelets.

Results: Twenty patients were included. Twelve patients had undergone BiMEx. All 20 patients had undergone SARME, of which 8 patients without MMD. Age at moment of surgery ranged from 16 to 47 years. There was a sagittal downward displacement of pogonion with tendency for increase of inter gonial distance. Furthermore, transversal expansion of nasal alar sulcus width and nasal base width was observed.

Conclusion: An automatic stereophotogrammetry landmarking analysis of soft tissue effects showed a sagittal downward displacement of pogonion following BiMEx and a transversal expansion of nasal alar sulcus width, and nasal base width after SARME.

4.1 Introduction

Transverse mandibular and maxillary deficiencies manifest in anterior and posterior crowding and/or in uni- or bilateral crossbite. Historically, these discrepancies were treated with orthodontic and/or dental extraction therapy. Since distraction was introduced for the facial skeleton in the early 90s of last century, new treatment options became possible [5, 15]

Mandibular midline distraction (MMD) is an effective technique to widen the mandible in order to solve transverse mandibular deficiencies [4, 5, 10].

For transverse maxillary deficiencies, surgically assisted rapid maxillary expansion (SARME) is an accepted technique and well reported in the literature [17][12][11] [13][18].

In some specific cases a combination of MMD and SARME is indicated, what is named as bimaxillary expansion (BiMEx)[3] [2]. Research on MMD is mostly performed using conventional research methods including dental cast models and posterior-anterior cephalograms [4], whereas for SARME outcome of studies using three-dimensional (3D) imaging analysis techniques is available [18]. However, research on BiMEx is reported sparsely in the literature up to now [1] [14], and to the authors' knowledge only one paper reports soft tissue effects following BiMEx using of 3D imaging analysis techniques[2].

Since 3D imaging techniques make it possible to analyse bony and overlying soft tissue structures more accurately compared with conventional two-dimensional (2D) radiographs, it is possible to obtain highly realistic skeletal and facial information. In addition, it is possible to acquire volumetric changes of bony and overlying soft tissue structures using 3D landmarking. This makes it possible to calculate

a prediction of facial changes following MMD and/or SARME.

Soft tissue effects could be evaluated by 3D facial surface scans or stereo photographs, and are obtained using stereophotogrammetry. The resulting data is a cloud of triangulated 3D points that forms a 3D model on which a full colour texture of the face can be mapped. 3D surface scans have been used in landmark-based clinical research [7] [6] with manually placed 3D landmarks. Recently, at the Erasmus University Medical Center, Rotterdam, the Netherlands a new method was created that can automatically place landmarks on facial surface data [9] [8].

The main objective of this study is to provide a 3D evaluation of the soft tissue effects following MMD and/or SARME.

4.2 Materials and methods

A retrospective observational study was conducted after approval had been given by the Medical Ethics Committee of Erasmus University Medical Center, Rotterdam, the Netherlands (approval number: MEC-2013-367).

4.2.1 Patients

From 2008 to 2013, patients who underwent MMD and/or SARME at the Department of Oral and Maxillofacial Surgery, Erasmus University Medical Center, Rotterdam, the Netherlands, were included in this study.

The inclusion criteria were mandibular discrepancy (mandibular anterior and/or posterior crowding, uni- or bilateral crossbite) treated with MMD, and maxillary discrepancy (maxillary anterior and/or posterior crowding and/or uni- or bilateral crossbite) treated with SARME. Patients were at least 16 years old.

The exclusion criteria were congenital (craniofacial) deformity patients, additional orthognathic surgery following MMD (bilateral sagittal split osteotomy) and SARME (Le Fort 1) before 1 year post-operative, mental retardation, history of radiation therapy and head injuries leading to fractures and/or soft tissue scars in the facial area of interest, missing stereophotogrammetry record at T1 and/or T3 and insufficient stereophotogrammetry record quality by artefacts or obstructing hair in the facial area of interest.

For MMD, the surgical technique was similar to the described technique of Mommaerts et al [16] and only bone-borne distractors were used [3]. For SARME, the surgical technique applied was described by Koudstaal et al [11] and only tooth-borne distractors (Hyrax) were used. For MMD and SARME both, the surgical intervention was performed under general anaesthesia. At fixed time points, stereophotogrammetry records were taken: pre-operative (T1), direct post-distraction (T2) and 1-year post-operative (T3).

4.2.2 Stereophotogrammetry analysis

A 3D stereophotogrammetry setup with 4 cameras (EOS 1000D, CANON INC.) and an integrated software (DI3Dcapture, Dimensional Imaging, Version 6.8.16.4255) were used to capture 3D photographs of the face. All photographs were taken with natural head position and relaxed facial musculature.

The stereophotogrammetry analyses were performed with an automatic 3D facial landmarking algorithm combining template with shape based methods as described elsewhere [9] [8]. In short, the automatic landmarking algorithm aligns the 3D surface scans, projects to 2D, and extracts 2D features that serve as input for multiple base 2D landmarking algorithms. These base algorithms are then combined using ensemble learning. After the landmarks are located, they are reverted back to 3D. Additionally, correlations between landmark coordinates in the training sample are used in a principal components (PCs) guided search. 26 landmarks were automatically placed. Additionally, all landmark positions were manually checked by three observers (AG, JPG and MAJ) and repositioned if necessary on the 3D (Fig. 1) and flat (Fig. 2) view.

The stereophotogrammetry analysis was divided in 2 regions. For MMD, these regions were the condylar process, gonion, mouth, lower lip and pogonion. For SARME, the tip of the nose, nasal alar sulcus, nasal base, philtrum and upper lip were used as regions.

To assess the effect of MMD on the soft tissue structures, the following relevant point to point landmark distances were digitally measured: 25-23, 25-22, 23-22, 21-23, 21-25, 17-22, 17-21, 14-12, 22-1, 22-4, 18-16, 17-20, 17-16, 17-18, 26-24, 1-25 and 4-23.

For the effect of SARME on the soft tissue structures, the following relevant point to point landmark distances were digitally measured: 26-24, 11-5, 10-6, 9-7, 15-19, 16-18, 14-20, 8-12, 1-12, 4-12, 8-5, 8-10, 1-13, 4-13, 1-14, 4-14, 1-19 and 4-15. The landmark distances between the left and right lateral canthus (4-1), and left and right medial canthus (3-2) were used as a control measurement.

4.2.3 Statistical Analysis

Two-sided paired Samples T-test and two-sided Wilcoxon Sign-Rank tests were used to assess differences between T1 and T3. A Bonferroni correction (BC) was applied to adjust p-values for the MMD outcome (adjusted significance level $p < 0.0026$) and for the SARME outcome (adjusted significance level $p < 0.0025$), separately.

4.3 Results

4.3.1 Patients

Twenty patients fulfilled the inclusion criteria. All of the 20 patients had undergone a SARME. Twelve of these patients had undergone a BiMEx.

The age at the time of surgery ranged from 16 to 47 years. See Table 1 for the patient characteristics.

All the patients completed the treatment and the obtained transversal expansion for correcting the transversal discrepancy was obtained. Eleven out of the 20 patients underwent additional orthognathic

surgery after 1 year follow-up. During MMD, only in 1 patient the bone-borne distractor caused a dehiscence in the buccal mucosa underneath the lower lip. This was transient and healed within 2 weeks by frequent flushing.

4.3.2 Stereophotogrammetry analysis

In Table 2, the complete results of the stereophotogrammetry analysis are described for MMD. For the distance between landmark 22-1, with the T-test, there was a significant difference in the scores for T1 (mean = 119.17, standard deviation (SD) = 8.20) and T3 (mean = 122.36, SD = 7.54) conditions; t(11) = -4.5196, p = 0.000873. Even after applying Bonferroni correction, this significance still holds. A Wilcoxon Signed-Rank test indicated that the median post-test ranks were statistically significantly different than the median pre-test ranks Z = -3.06 p = 0.002218. This comparison is still significant after applying the BC. For the distance between landmark 22-4, with the T-test, there was a significant difference in the scores for T1 (mean = 119.29, SD = 7.93) and T3 (mean = 122.41, SD = 7.35) conditions; t(11) = -4.2171, p = 0.001444 (significant after BC). A Wilcoxon Signed-Rank test indicated that the median post-test ranks were statistically significantly different than the median pre-test ranks Z = -2.82 p = 0.004742. However, after applying BC, this significance is lost. These outcomes indicate a downward displacement of the pogonion.

Regarding the inter gonial distance (25-23), there was an insignificant difference in the scores for T1 (mean = 110.71, SD = 11.63) and T3 (mean = 114.44, SD = 13.78) conditions; t(11) = -1.7243, p = 0.112611. This outcome indicates a tendency for an increase of the inter gonial distance when looking to the soft tissue structures in this region, but however not significant.In Table 3, the complete results of the stereophotogrammetry analysis are described for SARME. For the distance between landmark 11-5, with the T-test, there was a significant difference in the scores for T1 (mean = 34.93, SD = 2.99) and T3 (mean = 37.13, SD = 3.32) conditions; t(19) = -5.6009, p = 0.000011. Even after applying Bonferroni correction, this significance still holds. A Wilcoxon Signed-Rank test indicated that the median post-test ranks were statistically significantly different than the median pre-test ranks Z = -3.88 p = 0.000056 (significant after BC). For the distance between landmark 10-6, with the T-test, there was a significant difference in the scores for T1 (mean = 24.82, SD = 2.13) and T3 (mean = 26.59, SD = 2.92) conditions; t(19) = -3.0049, p = 0.003641 (not significant after BC). A Wilcoxon Signed-Rank test indicated that the median post-test ranks were statistically significantly different than the median pre-test ranks Z = -3.17 p = 0.000804 (significant after BC). These outcomes indicate a transversal widening of the nasal alar sulcus width and nasal base width.

4.4 Discussion

In this retrospective observational study, we looked at 3D evaluation of the soft tissue effects following MMD and/or SARME. Stereophotogrammetry records at T1 and T3 were analysed with an automatic 3D facial landmarking algorithm using 2D Gabor wavelets as described by De Jong et al [9] [8]. The results showed a downward displacement of the pogonion with a tendency for an increase of the inter gonial distance. Furthermore, a transversal widening of the nasal alar sulcus width and nasal base width

was observed.

Regarding MMD, these results are similar to what was described by Bianchi et al [2]. In their study, a forward and downward displacement of the chin was observed with a forward projection of the lower lip [2]. It should be noted that simultaneous SARME was performed in their study and in the present study. Regarding the downward displacement of the pogonion in the present study, we think this is the effect of the maxillary downward displacement following SARME. This theory is strongly supported by Xi et al [20], as they observed a skeletal downward displacement of the maxilla with a clockwise rotation of the mandible and inferior chin displacement after only SARME [20]. Therefore, this should be interpreted as a result of BiMEx instead of the MMD in the present study.

Furthermore, there was no significant displacement observed of the lower lip in the present study. It must be noted that differences in lip projection could be created by dental movements due orthodontic treatment, which is not a solitary effect of MMD. This makes comparison and analysis difficult.

There was a tendency for increase of the inter gonial distance when looking to the soft tissue structures in this region. This outcome is in concordance with De Gijt et al [4], as they observed a slight increase of the skeletal ramal angle (RA) at T3. In their study, a bone-borne distractor was applied as well and this increase was not significant with no difference of the skeletal RA in the long-term (6.5 years) follow-up [4]. However, this outcome could be strongly related to the type of distractor. Tooth-borne distractors practice their force on dentoalveolar level and theoretically would create more posterolateral widening compared to bone-borne distractors, which practice their force anteriorly on basal bone level only. Related to this, in the gonion region the soft tissue effects might be different dependent on the type of distractor. To our knowledge, no study has been conducted to compare the soft tissue effects of both distractor types following MMD.

Regarding SARME, similar soft tissue effects were observed by Nada et al[20]. In their study, an increase in the nasal volume and alar width was observed at 22 months post-SARME [19]. This outcome is an aesthetic effect of SARME for clinicians, which has to be taken into account when planning the orthognathic surgery. In the present study, there was a mean increase of 2.20 mm for the nasal alar sulcus width and a mean increase of 1.77 mm for the nasal base width. Although these increases are minimal, it is difficult to predict how the patients will experience these soft tissue effects from aesthetic aspects.

In the present study, a limitation is that the T2 stereophotogrammetry records were not complete for all the included patients. This made it impossible to analyse the soft tissue effects of MMD and/or SARME during the treatment at end of distraction. Since aesthetic aspects are getting more importance in the orthognathic surgery, it is essential to provide the patients a prediction of the soft tissue effects during the treatment as well. There was a downward displacement of the pogonion after BiMEx. However, this outcome does not provide a prediction of soft tissue effects for patients who will undergo MMD without simultaneous SARME. BiMEx seems to be beneficial for patients with a short lower third part of the face. On the other hand, BiMEx could lead to undesirable soft tissue effects for patients with a pre-existing gummy smile and long face. The transversal widening of the nasal alar sulcus width and nasal base width after SARME could be undesirable as well for patients. Clinicians should communicate these possible soft tissue effects with the patients carefully during the planning of the orthognathic surgery. The soft tissue effects of MMD without simultaneous SARME are not clarified yet. There is

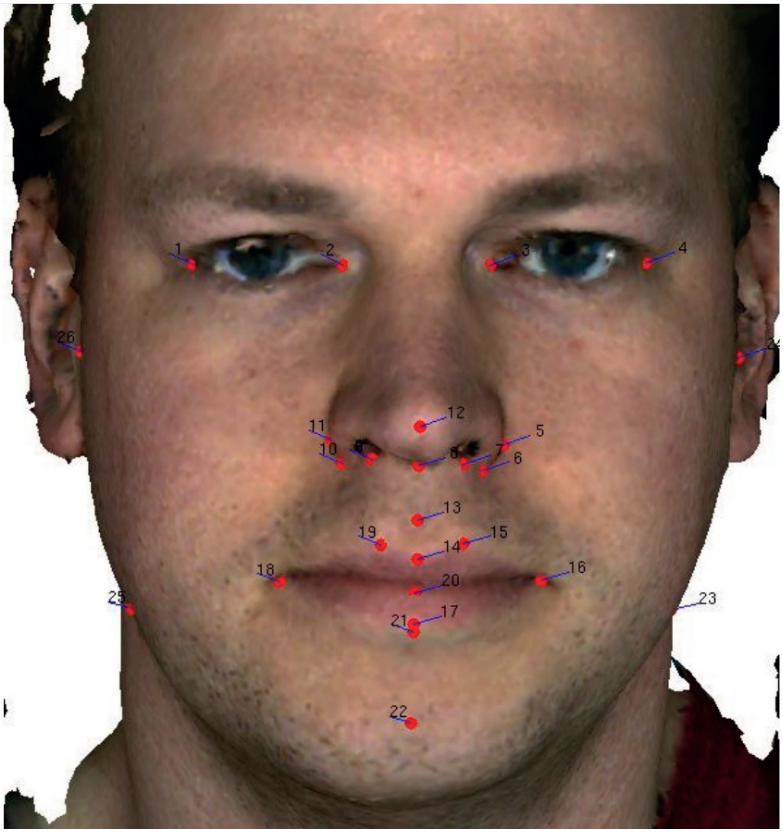


Figure 1: Overview of 26 automatically placed facial landmarks on the 3D view.

Figure 2: Overview of 26 automatically placed facial landmarks on the flat view.

still a lack of knowledge about the difference between the soft tissue effects of the different types of distractors following MMD.

Table 1. Baseline patient characteristics.

T1-T3	BiMEx	SARME (without MMD)	SARME (all)
Number of patients	12	8	20
Mean age (range)	29 (16-45)	31 (18-47)	30 (16-47)
Female:Male	8:4	5:3	13:7

Abbreviations: BiMEx, bimaxillary expansion; MMD, mandibular midline distraction; SARME, surgically assisted rapid maxillary expansion.

Table 2. Stereophotogrammetry analysis for MMD.

Landmark no.		T1 mean	T1 SD	T3 mean	T3 SD	diff	diff SD	sig	p-val	t	α	sig w	signed rank w	Z w	α w
1	4	90.23	3.30	90.75	3.70	0.52	1.52	0	11	-1.174614	0.264951	0	24	-1.176697	0.239317
2	3	35.67	2.36	36.03	2.42	0.36	1.27	0	11	-0.981601	0.347393	0	26	-1.019804	0.307821
25	23	110.71	11.63	114.44	13.78	3.73	7.50	0	11	-1.724262	0.112611	0	17	-1.725822	0.084379
25	22	88.66	9.37	89.99	9.47	1.33	5.50	0	11	-0.836595	0.420623	0	26	-1.019804	0.307821
23	22	89.79	11.03	93.43	11.88	3.64	6.85	0	11	-1.842050	0.092566	0	22	-1.333590	0.182338
21	23	90.80	10.27	94.08	11.73	3.28	6.78	0	11	-1.674721	0.122151	0	24	-1.176697	0.239317
21	25	89.35	8.68	90.38	8.96	1.03	5.79	0	11	-0.613754	0.551859	0	32	-0.549125	0.582920
17	22	34.18	5.54	35.12	4.69	0.95	5.05	0	11	-0.649258	0.529495	0	30	-0.706018	0.480177
17	21	22.17	4.90	22.02	4.96	-0.15	3.62	0	11	0.140256	0.890993	0	42	0.235339	0.813945
14	12	34.17	4.38	34.85	3.56	0.68	2.16	0	11	-1.091726	0.298288	0	25	-1.098250	0.272095
22	1	119.17	8.20	122.36	7.54	3.19	2.44	1	11	-4.519573	0.000873	1	0	-3.059412	0.002218
22	4	119.29	7.93	122.41	7.35	3.11	2.56	1	11	-4.217051	0.001444	1	3	-2.824072	0.004742
18	16	49.28	3.10	50.44	3.80	1.16	3.17	0	11	-1.266656	0.231442	0	24	-1.176697	0.239317
17	20	10.57	1.72	10.99	2.07	0.42	2.45	0	11	-0.588115	0.568333	0	33	-0.470679	0.637870
17	16	30.61	2.52	30.38	2.30	-0.24	2.07	0	11	0.397335	0.698722	0	40	0.078446	0.937473
17	18	30.39	2.22	30.13	3.51	-0.26	2.68	0	11	0.342600	0.738356	0	49	0.784465	0.432768
26	24	138.36	11.32	139.28	11.17	0.92	2.82	0	11	-1.125793	0.284216	0	26	-1.019804	0.307821
1	25	98.59	7.78	95.42	9.06	-3.17	6.48	0	11	1.691324	0.118876	0	58	1.490483	0.136097
4	23	99.00	8.34	95.40	8.75	-3.60	6.51	0	11	1.914878	0.081861	0	60	1.647376	0.099481

Abbreviations: α, alpha; diff, difference; MMD, mandibular midline distraction; p-val, *p*-value; SD, standard deviation; sig, significance; t, 2-sided paired Samples T-test; w, Wilcoxon Sign-Rank test.

Values are reported in millimeters.

2-sided paired Samples T-test result and 2-sided Wilcoxon Sign-Rank test results (Wilcoxon indicated with 'w' postfix).

Rows marked with + show left sided results (h1 = T1 > T3).

Note: The Bonferroni correction adjusts the α from 0.05 to 0.0026316.

Table 3. Stereophotogrammetry analysis for SARME.

Landmark no.		T1 mean	T1 SD	T3 mean	T3 SD	diff	diff SD	sig	p-val	t	α	sig w	signed rank w	Z w	α w
1	4	91.83	4.12	91.99	4.51	0.15	1.65	0	19	-0.412675	0.684464	0	93	-0.447992	0.654159
2	3	36.61	3.16	37.04	3.30	0.43	1.40	0	19	-1.374369	0.185323	0	65	-1.493307	0.135357
26	24	138.49	10.52	138.74	10.35	0.25	3.22	0	19	-0.344361	0.734357	0	82	-0.858651	0.390533
11	5	34.93	2.99	37.13	3.32	2.20	1.76	1	19	-5.600907	0.000011	1	1	-3.882598	0.000056+
10	6	24.82	2.13	26.59	2.92	1.77	2.63	1	19	-3.004860	0.003641	1	20	-3.173277	0.000804+
9	7	18.40	2.64	19.14	2.44	0.74	2.58	0	19	-1.285564	0.107022	0	88	-0.634655	0.268951+
15	19	13.43	2.41	13.46	2.15	0.02	1.87	0	19	-0.058476	0.953980	0	96	-0.335994	0.736875
16	18	49.25	3.80	50.12	4.03	0.87	3.67	0	19	-1.058566	0.303071	0	85	-0.746653	0.455273
14	20	8.34	1.73	8.40	2.59	0.06	2.00	0	19	-0.140136	0.890028	0	148	1.605305	0.108427
8	12	19.80	2.34	19.72	2.34	-0.08	2.84	0	19	0.130809	0.897302	0	98	-0.261329	0.793839
1	12	71.18	4.00	71.22	4.41	0.03	1.98	0	19	-0.077292	0.939200	0	123	0.671988	0.501591
4	12	71.47	4.26	71.18	4.38	-0.28	2.08	0	19	0.611588	0.548059	0	110	0.186663	0.851925
8	5	22.70	1.53	23.56	2.71	0.86	2.72	0	19	-1.412151	0.174075	0	84	-0.783986	0.433048
8	10	15.52	1.71	16.13	1.99	0.61	1.66	0	19	-1.655499	0.114247	0	69	-1.343976	0.178956
1	13	73.43	4.27	73.47	4.20	0.04	1.39	0	19	-0.119645	0.906020	0	118	0.485325	0.627446
4	13	74.13	4.25	74.02	4.40	-0.11	2.17	0	19	0.225185	0.824239	0	97	-0.298661	0.765198
1	14	81.34	4.30	81.96	4.48	0.62	1.81	0	19	-1.528066	0.142973	0	69	-1.343976	0.178956
4	14	82.00	4.20	82.53	4.26	0.54	1.76	0	19	-1.366501	0.187736	0	67	-1.418641	0.156004
1	19	75.91	4.17	76.52	4.11	0.61	1.96	0	19	-1.383187	0.182647	0	75	-1.119980	0.262722
4	15	76.54	4.04	77.11	4.06	0.56	1.74	0	19	-1.443236	0.165240	0	76	-1.082647	0.278965

Abbreviations: α, alpha; diff, difference; p-val, *p*-value; SARME, surgically assisted rapid maxillary expansion; SD, standard deviation; sig, significance; t, 2-sided paired Samples T-test; w, Wilcoxon Sign-Rank test.

Values are reported in millimeters.

2-sided paired Samples T-test result and 2-sided Wilcoxon Sign-Rank test results (Wilcoxon indicated with 'w' postfix).

Rows marked with + show left sided results (h1 = T1 > T3).

Automatic stereophotogrammetry landmarking analysis of soft tissue effects showed a downward displacement of the pogonion following BiMEx and a transversal widening of the nasal alar sulcus width, and nasal base width after SARME. Clinicians should communicate these possible soft tissue effects with the patients carefully during the planning of the orthognathic surgery.

References

[1] Mehmet Bayram et al. "Nonextraction treatment with rapid maxillary expansion and mandibular symphyseal distraction osteogenesis and vertical skeletal dimensions". In: *The Angle orthodontist* 77.2 (2007), pp. 266–272.

[2] Francesca Antonella Bianchi et al. "Soft, hard-tissues and pharyngeal airway volume changes following maxillomandibular transverse osteodistraction: Computed tomography and three-dimensional laser scanner evaluation". In: *Journal of Cranio-Maxillofacial Surgery* 45.1 (2017), pp. 47–55.

[3] JP De Gijt et al. "Mandibular midline distraction: a systematic review". In: *Journal of Cranio-Maxillofacial Surgery* 40.3 (2012), pp. 248–260.

[4] JP de Gijt et al. "Long-term (6.5 years) follow-up of mandibular midline distraction". In: *Journal of Cranio-Maxillofacial Surgery* 44.10 (2016), pp. 1576–1582.

[5] CA Guerrero et al. "Mandibular widening by intraoral distraction osteogenesis". In: *British Journal of Oral and Maxillofacial Surgery* 35.6 (1997), pp. 383–392.

[6] Peter Hammond et al. "Discriminating power of localized three-dimensional facial morphology". In: *The American Journal of Human Genetics* 77.6 (2005), pp. 999–1010.

[7] Peter Hammond et al. "Fine-grained facial phenotype–genotype analysis in Wolf–Hirschhorn syndrome". In: *European Journal of Human Genetics* 20.1 (2012), p. 33.

[8] Markus A Jong et al. "Ensemble landmarking of 3D facial surface scans". In: *Scientific Reports* 8.1 (2018), p. 12.

[9] Markus A de Jong et al. "An automatic 3d facial landmarking algorithm using 2d gabor wavelets". In: *IEEE Transactions on Image Processing* 25.2 (2016), pp. 580–588.

[10] John W King et al. "Long-term skeletal and dental stability of mandibular symphyseal distraction osteogenesis with a hybrid distractor". In: *American Journal of Orthodontics and Dentofacial Orthopedics* 141.1 (2012), pp. 60–70.

[11] MJ Koudstaal et al. "Stability, tipping and relapse of bone-borne versus tooth-borne surgically assisted rapid maxillary expansion; a prospective randomized patient trial". In: *International journal of oral and maxillofacial surgery* 38.4 (2009), pp. 308–315.

[12] MO Lagravere et al. "Dental and skeletal changes following surgically assisted rapid maxillary expansion". In: *International journal of oral and maxillofacial surgery* 35.6 (2006), pp. 481–487.

[13] Katharina Laudemann et al. "Long-term 3D cast model study: bone-borne vs. tooth-borne surgically assisted rapid maxillary expansion due to secondary variables". In: *Oral and maxillofacial surgery* 14.2 (2010), pp. 105–114.

[14] Sıddık Malkoç, Serdar Üşümez, and Haluk İşeri. "Long-term effects of symphyseal distraction and rapid maxillary expansion on pharyngeal airway dimensions, tongue, and hyoid position". In: *American Journal of Orthodontics and Dentofacial Orthopedics* 132.6 (2007), pp. 769–775.

[15] JG McCarthy et al. "Lengthening the human mandible by gradual distraction." In: (1992).

[16] MY Mommaerts. "Bone anchored intraoral device for transmandibular distraction". In: *British Journal of Oral and Maxillofacial Surgery* 39.1 (2001), pp. 8–12.

[17] MY Mommaerts. "Transpalatal distraction as a method of maxillary expansion". In: *British Journal of Oral and Maxillofacial Surgery* 37.4 (1999), pp. 268–272.

[18] Rania M Nada et al. "Three-dimensional prospective evaluation of tooth-borne and bone-borne surgically assisted rapid maxillary expansion". In: *Journal of Cranio-Maxillofacial Surgery* 40.8 (2012), pp. 757–762.

[19] Rania M Nada et al. "Volumetric changes of the nose and nasal airway 2 years after tooth-borne and bone-borne surgically assisted rapid maxillary expansion". In: *European journal of oral sciences* 121.5 (2013), pp. 450–456.

[20] Tong Xi et al. "The effects of surgically assisted rapid maxillary expansion (SARME) on the dental show and chin projection". In: *Journal of Cranio-Maxillofacial Surgery* 45.11 (2017), pp. 1835–1841.

5

Automated human skull landmarking with 2D Gabor Wavelets

Markus A. de Jong
Atilla Gül
Jan Pieter de Gijt
Maarten J. Koudstaal
Manfred Kayser
Eppo B. Wolvius
Stefan Böhringer

Abstract

Landmarking of CT scans is an important step in the alignment of skulls that is key in surgery planning, pre-/post-surgery comparisons, and morphometric studies. We present a novel method for automatically locating anatomical landmarks on the surface of cone beam CT-based image models of human skulls using 2D Gabor wavelets and ensemble learning. The algorithm is validated via human inter- and intra-rater comparisons on a set of 39 scans and a skull superimposition experiment with an established surgery planning software (Maxilim). Automatic landmarking results in an accuracy of 1-2 mm for a subset of landmarks around the nose area as compared to a gold standard derived from human raters. These landmarks are located in eye sockets and lower jaw, which is competitive with or surpasses inter-rater variability. The well-performing landmark subsets allow for the automation of skull superimposition in clinical applications. Our approach delivers accurate results, has modest training requirements (training set size of 30-40 items) and is generic, so that landmark sets can be easily expanded or modified to accommodate shifting landmark interests, which are important requirements for the landmarking of larger cohorts.

5.1 Introduction

An important clinical application of three-dimensional (3D) skull landmarking is skull morphometrics which requires dense correspondence or superimposition of pseudo-landmarks between pairs of skulls for analysis. Applications of skull morphometrics include bone growth analysis,[23] surgery planning and pre/post-surgery comparison and evaluation.[14] Usually, such comparisons between two skulls are initiated with a manual step in which a small set of anatomical landmarks (typically four or five) is placed on both skulls. Correspondence between the skulls for this set of landmarks is used to establish correspondence for a dense set of pseudo-landmarks by minimizing surface distances while respecting correspondence of anatomical landmarks by using, for example, thin plate splines. For instance, surgery planning software such as Maxilim uses this approach.[12] Pseudo-landmarks typically result from the set of points that define the surface of the skull and are derived from voxel data in the case of CT scans.

For larger landmark sets, a relevant and current application is facial genetics.[3,10,16] Other areas include facial reconstruction for both forensic[17] and archaeological purposes,[19] bone age determination[6] as well as sex determination[11] when DNA profiling is not available.

Manual landmarking of skull sets has several drawbacks. First, it is a tedious and time-consuming task for large skull cohorts. Second, non-trivial landmark definitions may introduce larger inter-rater disagreement. Third, it becomes costly to revise the set of anatomical landmarks. Fully automated and reliable landmarking of 3D human skull data therefore has important implications for morphometry-based research.

The body of literature on the topic of automated 3D human skull landmarking is small, even in comparison with literature on automated 3D facial landmarking. A non-voxel method exists,[25] but this method only attempts to find a small landmark set heuristically through a fitting with template skull model with known landmarks. Some voxel-based methods have also been published[4,15] that apply a registration to atlas images or template matching based on per-slice CT contours, respectively. Voxel-based methods have the advantage of using the full data set when voxel data is available but cannot be used when only surface data is available.[13] On the other hand, voxel data can be reduced to surface data.

Another approach that has previously been applied in facial data is the use of established 2D-based landmarking methods as in-between for 3D landmarking, such as active shape models.[26] While active shape model implementations for faces are common-place and freely available and ready-to-use on-line, no such resource exists for skull data. The creation of active shape models for skulls from scratch would typically require thousands of manually trained samples, when merely locating such large skull sets alone would already be difficult.

Here, we seek to develop a completely automatic landmarking algorithm that can process 3D models including both voxel and surface data and has low training complexity. We achieve this by transforming voxel data into surface data and landmarking this data. We use ideas from work in facial surface data[8] that involves a loss-less data transformation of 3D scans to layers of 2D data. From these, 2D features can be derived to define templates at the landmark sets in training data which can be used as input for learning parameters of the landmarking algorithm. After landmarks have been located in the feature

space, they can be reverted back to 3D.

The paper is organized as follows. We begin with a detailed description of the algorithm. Next, we test the algorithm with tomography-based data in the form of cone beam CT scans. We then validate these results through a comparison with manually placed, multi-rater landmarks. Here, base accuracy is evaluated by calculating Euclidean distance between automatically located landmark coordinates and human training landmark coordinates. Finally, we perform an experiment in a clinical application with Maxilim, a commercially available surgery planning software, and evaluate our findings.

5.2 Methods

Overview of the algorithm

For voxel input, we first create a 3D surface model of the data. Next, our algorithm applies a map projection to convert the 3D data into a 2D representation of texture, height above the projection surface, and curvature as derived in 3D space (main features). From these, a larger set of 2D features is generated by transforming these data components (derived features). A generic and accurate 2D landmarking method is applied to locate the landmarks using ensemble methods by combining landmark proposals from base landmarking algorithms.[7] A base landmarking algorithm can be any algorithm that can propose a landmark position given new input data. To abbreviate, we will refer to an individual landmarking algorithm as a *landmarker* in the following. Most base landmarkers perform a constrained template search based on Gabor 2D filters applied to the features. Finally, the 2D landmark coordinates are mapped back to 3D.

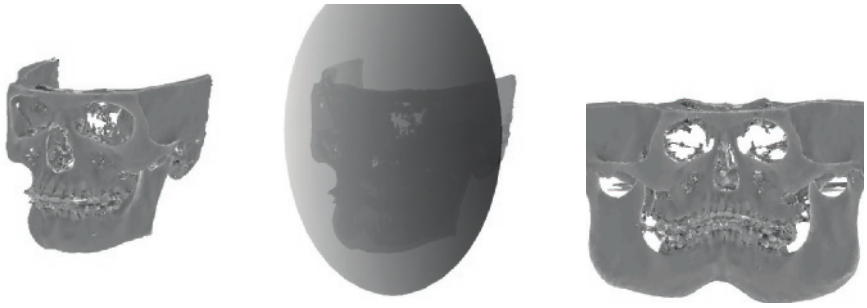
Data preparation

In this paper, we consider cone beam CT data available in the DICOM image file format. To prepare the data for landmarking, cone beam CT data is converted into high detail surface models using free open source software: the medical imaging and analysis software *3D Slicer*[5,21] and the 3D graphics software *Blender*.[2] A minimal cut-off Hounsfield value of 350 is used to isolate the bone data from the conic CT scans. After conversion to surface scans, the number of 3D data points (vertices) is ca. 600.000 with ca. 900.000 connecting edges. This data preparation process is automated into a single batch job that calls the required external programs.

In order to optimize landmarking results, the surface models are automatically aligned using an existing method[18] that combines a cylinder fitting approach with a 2D symmetry plane detection method that converges towards symmetry between the left and right hand sides of the frontal skull.

Map projection

The map projection process is illustrated in Figure 5.1. First, a standardized ellipsoid is fitted over the surface model (Figure 5.1b). Based on this ellipsoid, a Mercator map projection is performed on



(a) 3D skull surface model (b) Reference ellipsoid added (c) Map projected surface model

Figure 5.1: **Overview of the map projection process.** After the CT data is converted into a high detail surface model (a), a reference ellipsoid is fitted over the surface model (b) after which a Mercator map projection of the surface model is made based on the reference ellipsoid (c).

the skull data, converting it to 2D coordinates whilst retaining relative height information so that the resulting data set represents the map projected original surface model (Figure 5.1c).

2D Feature Generation

The 2D features that serve as input for the algorithm are derived from the map-projected model and are 200x200 pixels. A depth limit is applied on the map-projected model to reduce the influence of underlying bone structures that could add noise, *e.g.* thin and irregular bone structures visible through the eye sockets and nose.

The feature set for the landmarking algorithm consists of 4 main features from each of which 4 more edge enhancing transformations are derived, bringing the total number of features to 20.

Main feature 1, the Render feature, is a surface rendering performed with standardized lighting conditions. A uniform white surface color is used as in contrast to *e.g.* facial data, no texture information is available. Main feature 2, the Heigthmap feature, is a heightmap with respect to the fitted ellipsoid, where height is represented by a grey scale value. Main features 3 and 4, the Curvature features, use curvature information and represent the average curvature value by a grey scale value. Feature 3 uses the curvature information derived from the skull's un-projected state, 4 uses the curvature information derived from its projected state.

Based on these four main features, 4 sub-features are created with edge enhancement filters. These transformations are the derivatives over the x- and y-axis, a Laplacian of Gaussian (LoG) filter and a Sobel filter. An example of a complete feature set for one subject is given in Figure 5.2.

Base landmarking

We use the features defined above to implement several landmarking algorithms (base landmarkers) which are later combined into the full algorithm. Base landmarkers considered here are based on the Elastic Bunch Graph Matching method (EBGM)[24] which, in short, performs a maximum correlation

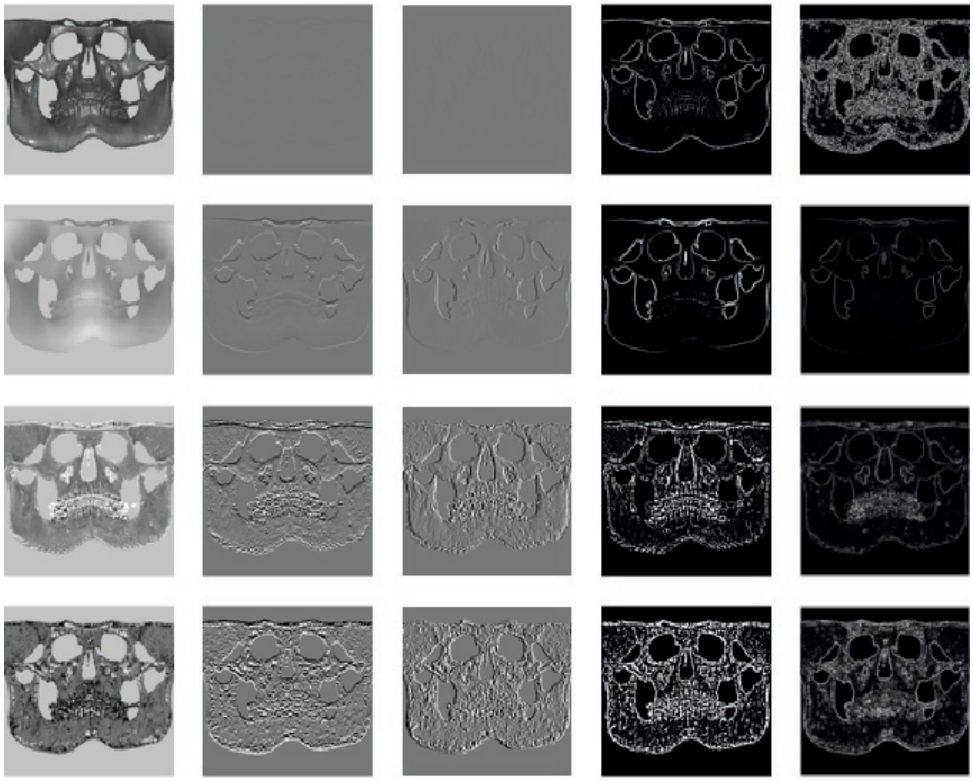


Figure 5.2: **Example 2D feature set for a single subject.**
Rows: 1) Render 2) Heightmap 3) Curvature projected model 4) Curvature original model.
Columns: a) Original b) derivative over y-axis c) derivative over X-axis d) Laplacian of Gaussian filter e) Sobel filter

template search between a set of example (training) images and the image to be landmarked restricted by the geometry of an average graph.

Templates for each feature are constructed by storing filter responses of 2D Gabor Wavelets in different sizes and rotations. The collection of all responses for all landmarks is also called the 'bunch graph'. Convolutions of the same 2D Gabor Wavelets with target images serve as input for the correlation search.

In the landmarking phase, correlations of wavelet responses at a set of candidate landmarks with subsets of the wavelet coefficients stored in the bunch graph are computed. Details of the process are described elsewhere.[8] Each feature serves as the input for a single landmarker. The output of each of the landmarkers is the coordinate indicated by the highest correlation of the target image with the template for each landmark.

Ensemble landmarking

The results of the individual landmarkers serve as input for a stacked generalization algorithm which automatically selects the best performing landmarkers for each landmark's x- and y-position and combines their results by means of linear regression. The linear regression is fitted on training data regressing the true landmark location on predictions by base landmarkers. Further details of the ensemble method are described elsewhere.[7]

As linear stacking does not constrain landmark predictions, it may cause some landmarks to be placed out of bounds of the projected skull model and end up leaving the support of the projected image. To avoid this issue, a local grid search is performed around predicted landmarks to find the nearest surface edge.

Landmark set

The 33 landmarks used in this paper are listed and described in Table 5.1 and illustrated in Figure 5.3. The landmarks are based on previous literature[20] and are amended by additional landmarks that might be relevant in certain applications. See Table5.1 for their anatomical descriptions.

Not all landmarks are available for each subject, e.g. some skulls are missing the upper part of the skull and with it landmark 6. We here do not consider the problem of missing landmarks and omit a landmark/subject combination from the evaluation when not available. These omissions are indicated in the results as total set-counts and are reported in the rightmost column of Table 5.1.

Training

The training phase consists of manually labeling the complete dataset of 39 subjects using a custom labeling software tool that allows the user to switch between the map projection and the original 3D model of the skull. The labels are placed on the map projected model after which the result can be inspected in 3D. Labeling typically takes circa 5 minutes per skull for 33 landmarks. The coordinates of the 2D model training landmarks are then translated into 2D feature coordinates to be used in the landmarking algorithm with sub-pixel accuracy.

5.3 Experiments

Data sets

The data set consists of 39 facial cone beam CT scans of patients acquired from the Oral and Maxillofacial Surgery and Special Dental Care department at Erasmus MC, Rotterdam, The Netherlands. The data set was derived from a non-syndromic cohort and anonymized. The slice thickness of the scans vary between .3mm and 1mm. The age of the subjects ranges from 16 to 54 with 28 as median and with 11 males and 28 females. The included cone-beam CT scans generally included the facial skeleton.

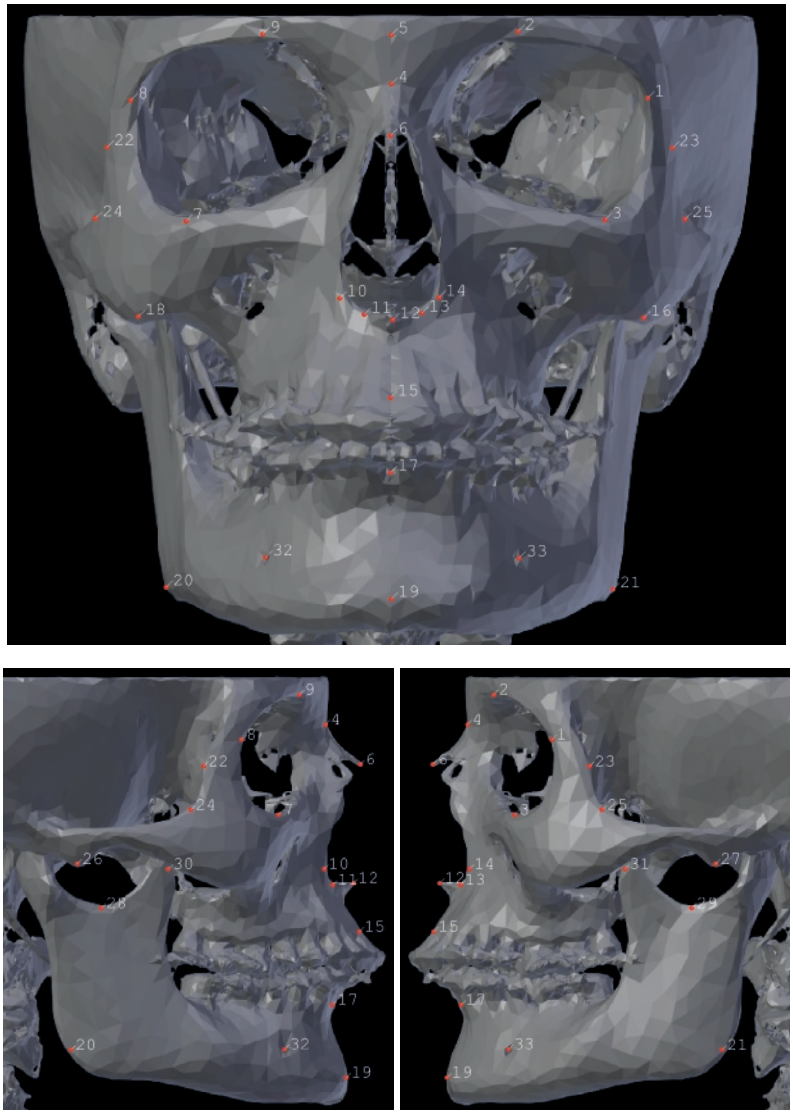


Figure 5.3: **Landmark set illustration.** The images show the landmark set from frontal, right and left view, respectively. Landmark descriptions are given in Table 5.1. The landmarks are shown on a for display purposes averaged, symmetrized, and coarsened skull model.

The facial scans ranged from the mandible to at least 1 cm above the orbits. Landmark availability count for this data set is given in the rightmost column of Table 5.1.

Data Availability

The 3D datasets generated and analysed during the current study are not publicly available due to patient privacy restrictions.

Cross-validation experiment

All 39 skulls were labeled by Rater 1 (MAdJ). Evaluation of landmarking accuracy was performed by automatically landmarking skulls in a 39-item leave-one-out setup, with the algorithm being trained on 38 samples and being applied to the left-out skull.

We have defined performance categories as follows, cased on distance comparisons with human raters and usability:[7]

- 1) good (<2mm): similar or better than human rater performance
- 2) medium (2-3mm): comparable to human rater for more "difficult" landmarks
- 3) poor (3-10mm): worse than human rating
- 4) extreme (> 10mm): unusable landmarks

Comparison with human raters

To perform an intra-rater comparison, Rater 1 (MAdJ) labeled the training set a second time using the custom labeling tool.

For an inter-rater comparison, a random subset of 29 skulls was manually labeled by two other raters (AG, JPdG) using the custom labeling tool.

For this experiment, the same performance categories apply as for the cross-validation experiment.

Field experiment with surgery planning software

An established surgery planning software, Maxilim,[12] uses manually placed landmarks to initialize the superimposition of skull models. In this experiment, we test the performance of several of our best performing landmark sub-sets to investigate whether our algorithm is ready to automate this kind of task.

As Maxilim does not allow for importing of landmark data, the landmarks were manually transferred from a laptop screen to the Maxilim input. Four landmarks are required to run the built-in hard-tissue superimposition procedure. These four landmarks must be placed on each skull at the same site. To get the best comparison of results and reduce superimposition uncertainty that could arise from intermediate growth etc., CT data from a single skull recorded at two instances during the same day were used for this experiment. Landmarks have been automatically placed on both recordings and two well-performing sub-sets of four landmarks (based on the results of the Cross-validation experiment) were subsequently used in two superimpositions.

5.4 Results

Landmarking results

Automatic landmarking Cross-validation results

Results of the 39 item leave-one-out setup are shown in Table 5.1 and illustrated in Figure 5.4. Results for algorithmic performance are given as Euclidean distances (given in *mm*) between automatically placed landmarks and manually placed landmarks (ground truth) in column *Algorithm*. Training was performed by rater 1 (MaDJ) and the ground truth were manual landmarks from the training set compared with automatically placed landmarks on the left-out sample. Algorithmic performance therefore measures the ability of the algorithm to mimic human rating. Rater 1 re-labeled the data set in an independent labeling session 10 weeks later for intra-rater comparison. Results are shown in column *Intra* as the mean distance between landmarks from the two labeling sessions. For inter-rater comparison, column *Inter* shows mean pair-wise distances between landmarks placed by raters MAdJ (ground truth), AG and JPdG.

ID	Description	Algorithm	Intra	Inter	Set size
1	Frontozygomatic (L)	2.0 (1.2)	**1.7** (1.4)	**1.9** (1.4)	36
2	Supraorbital foramen (L)	2.5 (2.1)	**1.4** (1.2)	2.2 (1.9)	26
3	Orbitale (L)	**1.9** (2.5)	**1.7** (1.6)	2.2 (2.7)	39
4	Nasion	3.0 (2.5)	**1.3** (1.7)	3.0 (3.2)	35
5	Transition nasal-frontal	2.7 (1.6)	2.1 (1.4)	2.7 (1.6)	18
6	Anterior nasal	**1.7** (1.4)	**1.2** (0.9)	**1.7** (1.3)	39
7	Orbitale (R)	3.7 (3.4)	**1.6** (1.1)	3.4 (3.2)	39
8	Frontozygomatic (R)	**1.5** (1.1)	2.4 (1.9)	**1.4** (1.8)	38
9	Supraorbital foramen (R)	2.9 (2.4)	3.0 (2.4)	2.4 (2.2)	27
10	Lateral nasal aperture (R)	5.6 (9.8)	2.4 (1.7)	6.3(12.3)	31
11	Distal nasal aperture (R)	**1.2** (0.8)	**0.9** (0.6)	**1.2** (0.8)	37
12	Anterior nasal spine	5.6 (8.1)	**1.1** (0.8)	6.2 (9.2)	36
13	Distal nasal aperture distal (L)	2.2 (4.4)	**1.3** (1.1)	2.6 (4.9)	35
14	Lateral nasal aperture (L)	2.8 (4.1)	2.5 (1.7)	3.1 (4.9)	30
15	Maxillary central incisors	3.7 (4.0)	**0.6** (0.6)	4.1 (4.3)	39
16	Maxilla-zuggomatic transition (L)	4.9 (3.8)	7.2(17.5)	4.3(16.8)	37
17	Mandibular central incisors	3.7 (3.0)	2.1 (3.5)	3.7 (4.0)	38
18	Maxilla-zuggomatic transition (R)	4.5 (3.1)	5.9(17.4)	3.9(16.5)	37
19	Anterior mental protuberance	4.6 (8.4)	**2.0** (1.1)	5.0 (9.5)	38
20	Gonion (R)	4.8 (5.7)	**1.9** (1.5)	3.4 (4.1)	38
21	Gonion (L)	2.6 (2.0)	2.0 (1.1)	2.4 (2.0)	39
22	Lateral zygomatic (R)	**1.7** (1.0)	**1.1** (2.0)	**1.8** (1.6)	39
23	Lateral zygomatic (L)	2.1 (1.1)	**1.4** (1.7)	2.1 (1.8)	39
24	Corner zygomatic (R)	**1.2** (0.6)	**1.3** (0.7)	**1.2** (0.7)	39
25	Corner zygomatic (L)	**1.1** (0.7)	**1.3** (0.6)	**1.0** (0.7)	39
26	Lateral zygomatic	2.6 (2.5)	**1.5** (1.0)	2.7 (2.7)	38
27	Zygomatic process	14.4 (31.1)	**1.7** (1.2)	8.9(20.9)	39
28	Distal mandibular notch (R)	8.8 (7.0)	**1.0** (0.6)	8.9 (7.1)	39
29	Distal mandibular notch (L)	13.7 (29.0)	**1.0** (0.7)	12.7(27.1)	39
30	Coronoid approximated (R)	8.0 (7.9)	**1.7** (4.5)	8.0 (9.4)	38
31	Coronoid approximated (L)	5.2 (7.0)	**0.7** (0.4)	5.2 (7.5)	38
32	Mental foramen (R)	**1.0** (0.8)	**0.8** (1.0)	**1.0** (1.0)	38
33	Mental foramen (L)	**1.4** (1.8)	**1.2** (2.2)	**1.5** (2.7)	39

Table 5.1: **Cross-validation results** of the set of 33 landmarks, given in Euclidean distances in millimeters to manual training data by Rater 1, standard deviations are given in parentheses, good performance in bold. *Algorithm*: automatic landmarking result distances to manual training data by Rater 1. *Intra*: Intra-rater: distances between first manual labeling and re-labeling by Rater 1. *Inter*: Inter-rater: distances to mean of manual Raters 1-3. *Set size*: training set size used for this landmark in the algorithm.

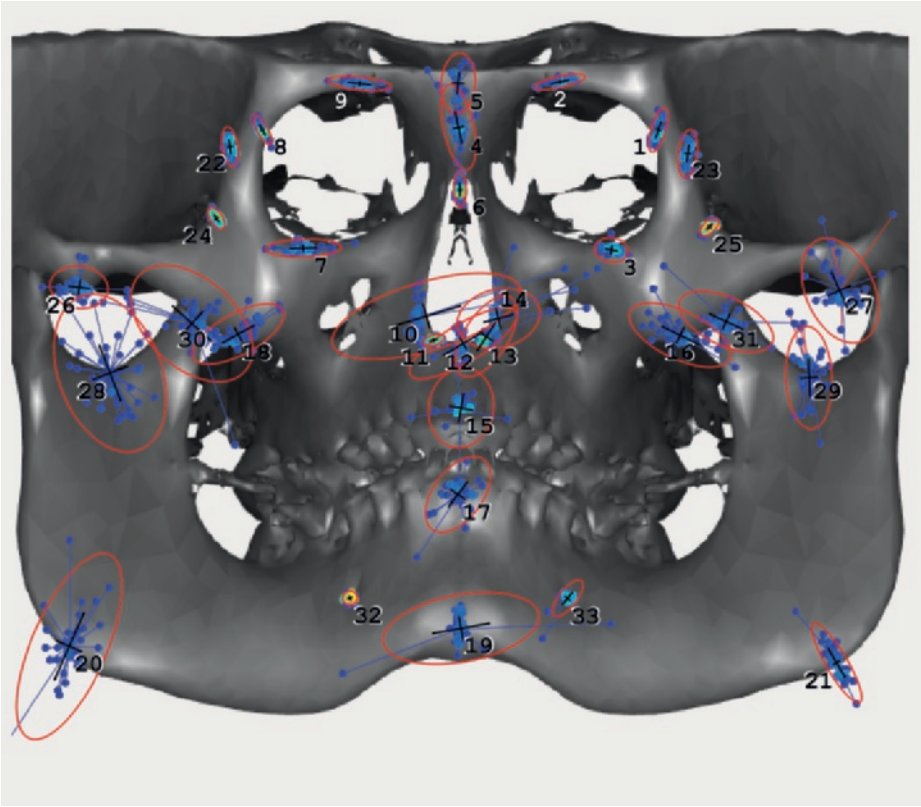
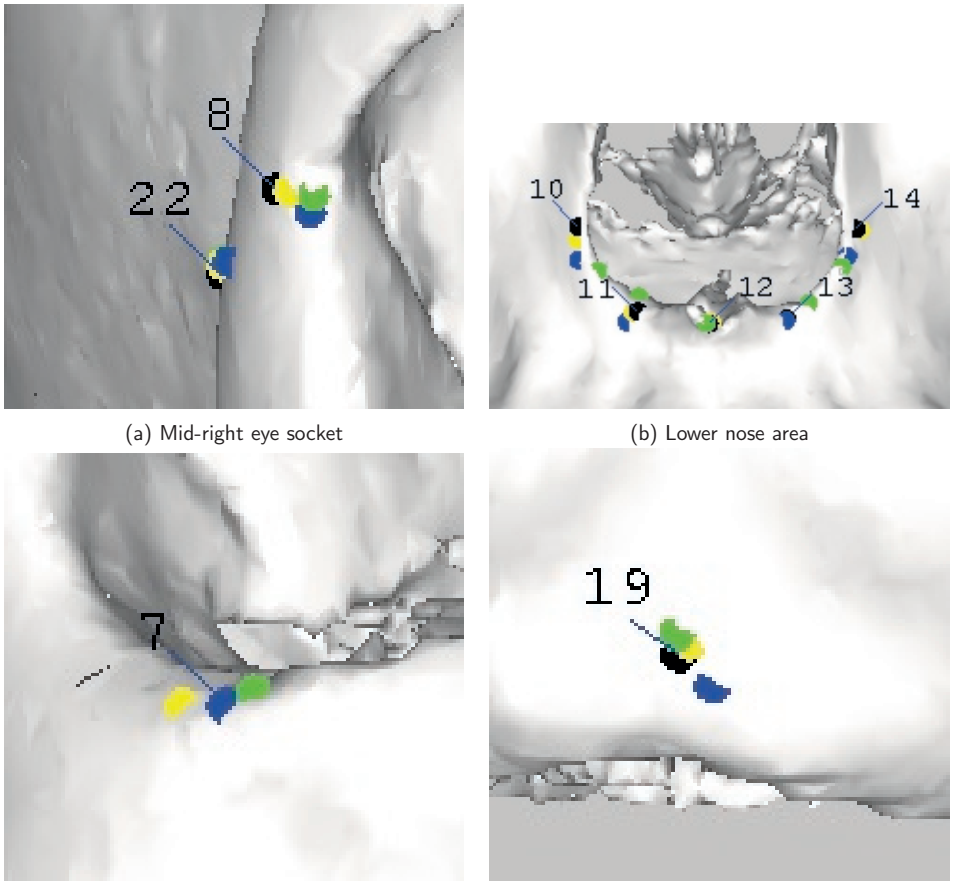


Figure 5.4: **Illustration of automatic 2D landmarking results for 33 landmarks and 39 subjects.**. For illustration purposes, all landmark results are plotted relatively to the ground truth of a projected skull model but are unrelated to this model. The red circles indicate result spread and direction.

When inspecting algorithmic performance in Table 5.1 (column *Algorithm*), 10 out of 33 landmarks perform well. Best performing landmarks are the two mental foramina (32, 33) and both corners of the zygomatic (24, 25). The remaining well-performing landmarks are both frontozygomatic landmarks (1, 8), orbitale (3), most anterior point of nasal bone (6), most distal point Nasal aperture (11) and the most lateral point zygomatic (22). Medium performing landmarks include landmarks that are symmetrical to well-performing landmarks (13, 23) and five more (2, 4, 5, 9, 14, 26). The remaining landmarks are landmarked unreliably with poor performance (10, 12, 28, 30, 31). Extreme performance are seen for two landmarks (27, 29).

Intra-rater results

Examples of intra-rater labelings are shown in Figure 5.5 (landmark colors black, yellow). Intra-rater results are displayed in Table 5.1 in column *Intra*. Overall, most intra-rater results (24 of 33) lie in the well-performing range, with 4 of those $<$1mm. Some landmarks show medium performances, however, such as landmarks nasion (4), frontozygomatic (8), supraorbital foramen (9), both lateral nasal aperture landmarks (10, 14) and the mandibular central incisors landmark (17). Poor performance can be seen



(a) Mid-right eye socket

(b) Lower nose area

(c) Lower right eye socket

(d) Chin

Figure 5.5: **Examples of rater variance in landmark placement.** See Table 5.1 for landmark descriptions.

Black: Rater 1 (training data, ground truth)

Yellow: Rater 1 (re-labeling for intra-rater comparison)

Green: Rater 2

Blue: Rater 3.

for both distal maxilla-zuggomatic transition landmarks (16, 18).

Inter-rater results

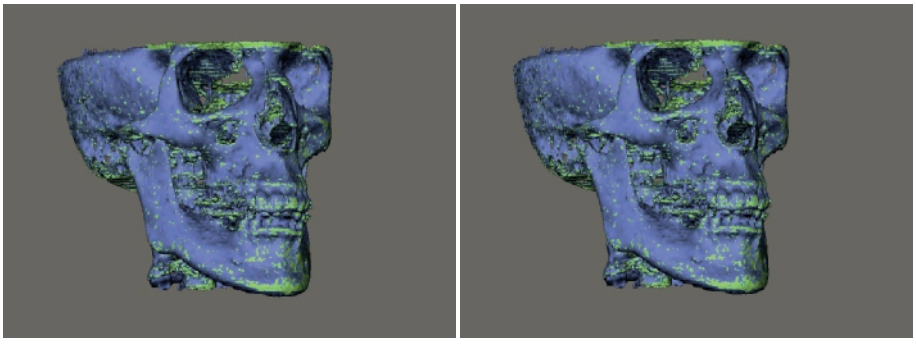
Sample inter-rater labelings are shown in Figure 5.5. The Figure clearly illustrates the presence of inter-rater disagreement.

Results of the inter-rater comparison are shown in Table 5.1 in column *Inter*. When inspecting the inter-rater distances, landmarks that perform well in relation to the mean of the three raters are both frontozygomatic (1, 8), anterior nasal (6), distal nasal aperture (11), lateral zygomatic (22), both corner zygomatic (24, 25) and both metal foramen landmarks (32, 33). The remaining landmarks show medium or worse performance with poor performance for the lateral nasal aperture (10), anterior nasal spine (12), anterior mental protuberance (19), zygomatic process (27), the first distal mandibular notch (28) and both coronoid approximated (30, 31) landmarks. One extreme score is reported for the second landmark on the distal point of mandibular notch (29).

Field experiment results

The superimposition results of two subsets of landmarks are displayed in Figure 5.6.

The following unique landmark subsets were selected based on performance to best illustrate our algorithms current potential:
Set 1: landmarks 6, 11, 24, 25
Set 2: landmarks 6, 11, 3, 8



(a) Landmark Set 1: 6, 11, 24, 25 (b) Landmark Set 2: 6, 11, 3, 8

Figure 5.6: **Skull superimposition results of the Maxilim field experiment based on 2 subsets of well-performing landmarks.** The two superimposed CT images were recorded on the same day and are displayed in blue and green, respectively.

5.5 Discussion

We present an automatic approach for landmarking of human skulls using a surface approach. One important clinical application is skull superimposition. We have demonstrated by manually transferring automatically placed landmarks into a commercial superimposition software that accurate, completely automatic skull superimposition is possible. The manual step was necessary for the purely technical reason of a lack of a means to import landmark coordinates into Maxilim. Such a feature can be easily implemented and we expect it to appear in the software as soon as automatic superimposition is demanded by users.

Due to lack of a golden standard, we evaluate accuracy by comparing automatic landmarking accuracy with respect to a single rater with intra- and inter-rater variation. Our results indicate that nine landmarks perform with absolute average discrepancy between automatically placed and training landmark of below 2mm. As the algorithm is limited by the accuracy of the training data a meaningful comparison can be made with intra-rater accuracy, *i.e.* how does algorithmic accuracy compare to consistency of a rater. This accuracy is comparable for the nine landmarks above, indicating that the algorithm can mimic a rater as well as the rater can mimic himself.

The inter-rater comparison of the distance between 3 raters and their mean also reports nine landmarks scoring $<$2mm. With higher standard deviations between raters than for the algorithm, this result implies higher between-rater difference as compared to automatic labeling. Should we be able to use a golden standard for landmark locations, our experiments suggest a better performance of automated landmarking over manual labeling.

Accuracy for individual landmarks usually agreed between automatic and manual labeling. When we compare standard deviations, we see that the standard deviation is generally lower in automatically placed landmarks for landmarks that perform better.

Using different subsets of well-performing landmarks, superimpositions perform well and show visually identical results that correspond to clinical expectations. This experiment illustrates that our algorithm can already be used to perform this task automatically and accurately.

When taking the nature of the landmarks into consideration, the inter-rater performance for some non-literature or experimental landmarks possibly suffers from ambiguous labeling instructions (landmarks 16, 18). In these cases the anatomical definition is difficult to reproduce on the skull. Here, the automatic landmarking algorithm shows a clear advantage. In contrast, human raters perform better at for example the gonion landmarks (20, 21) that rely on a deeper understanding of the anatomy which helps a human rater to intuitively find the best position. The gonion landmarks are not pronounced when visually inspecting the 2D features used in this study. Contrariwise, a human rater can switch back and forth between 2D to 3D in our labeling software to determine their best landmark position.

When looking at training set size, landmarks with less than the recommended 30 valid subjects[8] available for training (landmarks 2, 5 and 9) will benefit from a larger training set.

Furthermore, for symmetrical landmark pairs that are present on each side of the skull, we expect comparable performance (landmarks 8, 13 and 23). Observed differences in accuracy are in the range of 0.5-1mm. These differences can be explained by the different training sets and asymmetries in training

and test data as well as labeling errors. Taken together these findings highlight the importance of accurate training data. One exception is the pair of orbitale landmarks (3, 7) for which accuracies are 1.7 and 3.7mm, respectively. The errors for the right orbitale are mainly driven by the x-direction for which the landmark does have little structural information.

It is interesting to compare skull with facial surface data.[8] An important difference is that the bone structures of the skull model lead to far more complex models in terms of structural (dis)continuities and curvature. More so than in facial scans, some features such as nerve openings (foramen) or teeth may be missing in a subset of individuals. Furthermore, the cavernous nature of the skull may lead to more extreme outliers as a misplaced landmark in continuous facial data will simply end up a millimetre away any horizontal or vertical direction, while with skull data it might end up inside a cavity (e.g. eye socked), adding height distance.

A relevant limitation of our approach concerns landmark occlusion. This problem is shared among projection based methods, such as for facial surface scans. For example, landmark 31 can only be seen as an approximation of the coronoid as the actual coronoid is often hidden from view by the cheekbone in our current map projection set-up and the algorithm will choose a closest visible proxy in these cases. Tweaking the map projection and/or rotation of the 3D model or using multiple rotations could lead to better localization in such cases. Our current algorithm, however, does not contain modifications for specific anatomical structures. An advantage of skull data are its high resolution and sharp edges. This potentially allows for the expansion of the feature space. For example, multi-level resolution 2D features being smoothed at different levels of detail could could be used.

An important issue specific to CT data is the presence of artifacts, the largest contributor in our data being scatter caused by metal dental restorations that result in blobs and spikes, a reason why no landmarks were tested in those areas. Another factor limiting accuracy is the general quality of the image, where low-resolution CT images tend to result in larger cavities, e.g. in the zygomatico-maxillary suture area (the cheek bone). Finally, a specific limitation in cone beam data is the horizon of detail that results in fading detail towards the lateral side of the skull. In our application, this effect is not relevant as most of our landmarks are located close enough to the medial center of the skull.

The algorithm presented here, is potentially useful for a wide range of data types from which surface data can be derived. While this paper describes an automatic landmarking method for cone beam CT data, it is also compatible with other sources such as any 3D (tomographic) data and could be modified to work with, for instance any type of CT or MRI data as well of any type of surface data. As, for example, 3D data as acquired through photogrammetry is markedly cheaper than other sources (CT, MRI) and radiation free, our algorithm is potential useful in low-cost and low-radiation applications. Hand-held photogrammetrical recordings, for example, are becoming of interest in medicine,[1] forensics[22] and archaeology.[9]

In conclusion, our algorithm shows robust performance for 9 out of 33 landmarks with an accuracy of $<$2mm. This is comparable with the best performance in a similar algorithm for human faces,[7] and markedly better than another automated algorithm for human skulls (3.6-5.6 mm).[25] The low training burden of 30-40 subjects taking few minutes each and the non-heuristic nature of the algorithm allows landmark sets to be easily expanded or retrained to accommodate shifting landmark interests and heterogeneous samples. These results satisfy important requirements for the landmarking of large

(medical) cohorts. Even though the algorithm needs more work to support larger landmark sets, our results show that automatic skull superimposition is already feasible with the current iteration of our algorithm.

References

[1] Caroline AA Beaumont et al. "Three-dimensional surface scanners compared with standard anthropometric measurements for head shape". In: *Journal of Cranio-Maxillofacial Surgery* 45.6 (2017), pp. 921–927.

[2] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Institute, Amsterdam: Blender Foundation, 2016. url: `http://www.blender.org`.

[3] Stefan Boehringer et al. "Genetic determination of human facial morphology: links between cleft-lips and normal variation". In: *European Journal of Human Genetics* 19.11 (2011), pp. 1192–1197. issn: 1018-4813.

[4] Stefan Boehringer et al. "Genetic determination of human facial morphology: links between cleft-lips and normal variation". In: *European Journal of Human Genetics* 19.11 (2011), pp. 1192–1197.

[5] Andriy Fedorov et al. "3D Slicer as an image computing platform for the Quantitative Imaging Network". In: *Magnetic resonance imaging* 30.9 (2012), pp. 1323–1341.

[6] Elke Hillewig et al. "Magnetic resonance imaging of the medial extremity of the clavicle in forensic bone age determination: a new four-minute approach". In: *European radiology* 21.4 (2011), pp. 757–767.

[7] Markus A Jong et al. "Ensemble landmarking of 3D facial surface scans". In: *Scientific reports* 8.1 (2018), p. 12.

[8] Markus A de Jong et al. "An automatic 3D facial landmarking algorithm using 2D Gabor wavelets". In: *IEEE Transactions on Image Processing* 25.2 (2016), pp. 580–588.

[9] MIKOLÁŠ Jurda and PETRA Urbanova. "Three-dimensional documentation of Dolni Vestonice skeletal remains: can photogrammetry substitute laser scanning?" In: *Anthropologie* 54.2 (2016), p. 109.

[10] Fan Liu et al. "A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans". In: *PLoS Genetics* 8.9 (2012), e1002932. issn: 1553-7404.

[11] Li Luo et al. "Automatic sex determination of skulls based on a statistical shape model". In: *Computational and mathematical methods in medicine* 2013 (2013).

[12] Medicim NV. *Maxilim Maxillofacial Surgery Planning Software*. 2016. url: `http://www.medicim.com`.

[13] Morphosource Community. *www.morphosource.org*. Durham, North Carolina: Duke University, 2018. url: `http://www.morphosource.org`.

[14] Antonio Carlos de Oliveira Ruellas et al. "3D mandibular superimposition: comparison of regions of reference for voxel-based registration". In: *PloS one* 11.6 (2016), e0157625.

[15] Roshan N Rajapakse et al. "Automated Extraction of Cranial Landmarks from Computed Tomography Data using a Combined Method of Knowledge and Pattern Based Approaches". In: *Applied Medical Informatics* 38.1 (2016), p. 1.

[16] John R Shaffer et al. "Genome-wide association study reveals multiple loci influencing normal human facial morphology". In: *PLoS genetics* 12.8 (2016), e1006149.

[17] Laura J Short et al. "Validation of a computer modelled forensic facial reconstruction technique using CT data from live subjects: a pilot study". In: *Forensic science international* 237 (2014), 147–e1.

[18] L Spreeuwers. "Fast and Accurate 3D Face Recognition Using Registration to an Intrinsic Coordinate System and Fusion of Multiple Region". In: *Proc of Int Journal of Computer Vision* 93.3 (2011), pp. 389–414.

[19] Douglas R Stenton et al. "Faces from the Franklin expedition? Craniofacial reconstructions of two members of the 1845 northwest passage expedition". In: *Polar Record* 52.1 (2016), pp. 76–81.

[20] Gwen RJ Swennen. "3-D cephalometric soft tissue landmarks". In: *Three-dimensional cephalometry*. Springer, 2006, pp. 183–226.

[21] The Slicer Community. *3D Slicer - A multi-platform, free and open source software package for visualization and medical image computing*. 2016. url: `http://www.slicer.org`.

[22] Petra Urbanová, Petr Hejna, and Mikolaš Jurda. "Testing photogrammetry-based techniques for three-dimensional surface documentation in forensic pathology". In: *Forensic science international* 250 (2015), pp. 77–86.

[23] A Weissheimer et al. "Fast three-dimensional superimposition of cone beam computed tomography for orthopaedics and orthognathic surgery evaluation". In: *International journal of oral and maxillofacial surgery* 44.9 (2015), pp. 1188–1196.

[24] Laurenz Wiskott et al. "Face recognition by elastic bunch graph matching". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19.7 (1997), pp. 775–779. url: `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=598235` (visited on 06/05/2014).

[25] Kun Zhang, Yuan Cheng, and Wee Kheng Leow. "Dense correspondence of skull models by automatic detection of anatomical landmarks". In: *International Conference on Computer Analysis of Images and Patterns*. Springer. 2013, pp. 229–236.

[26] Dianle Zhou, Dijana Petrovska-Delacrétaz, and Bernadette Dorizzi. "Automatic landmark location with a combined active shape model". In: *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*. IEEE. 2009, pp. 1–7.

6

Automated asymmetry estimation of facial 3D scans using guaranteed symmetrical correspondence

Markus A. de Jong
Maarten J. Koudstaal
Pirro Hysi
Tim Spector
Eppo B. Wolvius
Manfred Kayser
Stefan Böhringer

Abstract

Asymmetry estimation is important for studies of human perception, anatomy, and surgery planning. Existing approaches start by registering a dense set of landmarks to the face and subsequently identifying corresponding landmarks in the two halves of the face. We consider a novel approach for automatic asymmetry registration by starting with a reference face, or atlas, that already has a well-defined symmetrical correspondence between nodes. This atlas, a synthetic 3D facial surface model generated out of a set of faces, is used to register each target face by bending it into the target shape. This procedure guarantees that corresponding landmarks are directly known for the target face. This process is supported by a set of 21 3D landmarks that are registered by an automatic landmarking procedure. To asses quality, we performed several simulation experiments by measuring symmetry registration effectiveness on controlled facial deformations, both point simulations as sagittal bending simulations. The simulation studies imply that asymmetries are correctly characterized by the algorithm.

6.1 Introduction

Facial symmetry is an important aspect in the perception of human faces and the perception of human attractiveness, the latter generally reflected by highly but not perfectly symmetrical faces. Although this relationship between attractiveness and asymmetry is difficult to quantify exactly [ref, ref], the opposite extreme of high asymmetry is perceived as dysmorphic and is often a result of genetic abnormality.[11]

By making use of 3D photogrammetry, the planning and evaluation of surgery aimed at improving facial symmetry can be assisted by accurate and objective evaluation of facial asymmetry. A different, more technical aspect of interest is the investigation of asymmetry evaluation itself, as it is not trivial to define appropriate measures of accuracy for these applications.

Assessment of facial symmetry appears to be intuitively and instantaneous for a person, albeit perhaps semi-unconsciously. However, attempting to quantitatively measure asymmetry is not. The problem of asymmetry estimation can be translated into the problem of finding corresponding landmarks on both sides of the face. From such a set of corresponding landmarks, asymmetry measures can be computed. However, this correspondence is unknown for a given 3D surface scan of a face and must be established by a registration process.

One approach is to mirror the scan with respect to the x-coordinate (horizontal axis), and overlaying the mirror scan with the original. Then, landmarks from both scans that are in close proximity to each are found and compared.[3,10] A major disadvantage of such a mirror-based approach is its heuristic nature, *i.e.* it is unclear what the precise relationship between original and mirrored landmarks is. For example, in faces with stronger asymmetries, left-right corresponding landmarks do not have to be close to each other after mirroring. Furthermore, these algorithms rely on the propagation of asymmetry throughout the face by smoothing methods such as thin plate splines which might exaggerate this problem. Also, initial face alignment required before mirroring might depend on the unknown asymmetry structure of the face and can only compare left and right over a straight line. Another disadvantage is that accuracy evaluation with a ground truth is difficult for such methods.

Atlas-based methods have been successfully applied in image registration.[6] In this application, pixels in the atlas are annotated by a segment class, such as tissue type, and after registering a target image to the atlas these annotations can be transferred. We apply atlassing to the asymmetry estimation by annotating left-right correspondence in the atlas and transferring this annotation to the target face by means of registration. The previously mentioned methods start with known landmark positions and try to establish correspondence. Atlassing starts with known correspondence and establishes landmark positions by means of registration, turning around the process. A priori, we see a number of advantages of this approach. Firstly, a synthetic 3D surface scan with perfect symmetry can be easily generated from a set of scans without any training. Secondly, by turning asymmetry estimation into a registration problem, previous work in automatic landmarking can be used. Thirdly, an atlas-based approach can arguably better deal with extremely asymmetrical faces as the registration process can be optimized for such cases. Finally, this approach standardizes the number of landmarks registered as all registered faces will contain the *atlas* landmarks, which is desirable in certain applications.

The aim of this study is to establish left-right correspondence using an atlas-based approach on a dense set of landmarks. Once this correspondence is established, individual, pair-wise asymmetries can be summarized into scores by various means which lie outside the scope of this paper.

This paper is organized as follows: we first describe our symmetry registration method in detail and illustrate the process with and example. In the next section we perform controlled symmetry experiments intended to measure the accuracy of the registration. We do this by applying a series of point deformations. Secondly, we perform experiments with the bending of the sagittal plane of the face, which is of interest as this is common in several facial syndromes. In the final section we discuss our results.

6.2 Methods

6.2.1 Input data

The input consists of unprocessed 3D photogrammetrically created 3D images consisting of 3D data points (vertices) and their connecting edges (vertexes) that together form a surface. On this surface, photographic data is projected (see Figure 6.1). The data was captured with a 3DMD camera system.[1]

Figure 6.1: Map projection process Left: 3D data points, Middle: connecting edges, Right: resulting surface with projected photographic data. The face used in this illustration is that of author MAdJ.

6.2.2 Data pre-processing

Before beginning to locate the 21 final landmarks, each face is first aligned in all 3 dimensions using an automated alignment tool. This tool uses a method that finds the unique mean curvature of the nose tip and both inner eye corners.[9] We then locate a point on the face between the eye corners. Through this middle point a line is drawn downwards over the surface of the face through the tip of the nose. We then analyse the slope of this ridge line to locate two valleys that represent one point below and one point above the nose. These two points serve as the basis for the final facial alignment of the dataset.[8]

6.2.3 Facial landmarking

After alignment, the 3D faces are map projected to 2D with a reference ellipsoid serving as a 'globe'. This transformation is reversible and retains the 2D coordinates' height information in relation to the reference ellipsoid. The map projection process is illustrated in Figure 6.2.

Figure 6.2: **Map projection process** of the 3D data (left). Middle: with a fitted reference ellipsoid serving as 'globe', the 3D shape is projected to 2D whilst retaining height information. Right: the green square illustrates the area of interest for automated 2D facial landmarking. The face used in this illustration is that of author MAdJ.

This process can be compared to the creation of a 3D relief map of a mountainous area. This map projection serves as basis for the input of an established and well-performing 2D landmarking method based on Gabor Wavelets. In short, we generate many 2D features that serve as input for the 2D landmarking algorithm. The set of 2D input-features is created by using different modalities of available information to colorize the map projected model (*i.e.* photographic information, height, curvature) and from these, create derivatives (*i.e.* a variety of edge enhancements). Each of these features are illustrated in Figure 6.3.

Each feature serve as the input for a landmarking algorithm or *landmarker*. The results for these landmarkers (and some combinations of intermediate, un-averaged results) form an ensemble of landmarkers on with which we apply stacked generalization to predict the best combination of inputs for each landmark.[5] This way we achieve the optimal 2D landmark result, illustrated in Figure 6.4]. These 2D landmarks are then map projected back to 3D. In this application, we use a set of 21 anatomical landmarks that is described elsewhere.[5]

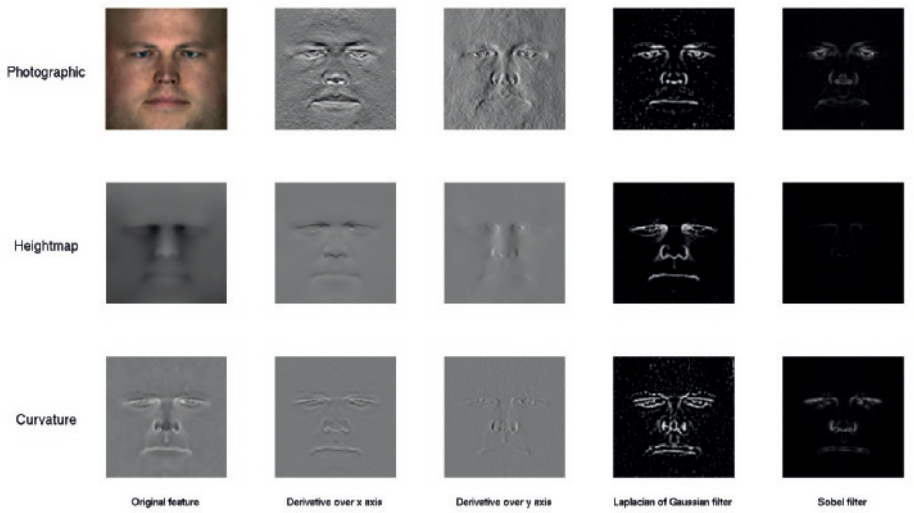


Figure 6.3: **Feature overview.** Each row represents a main information source: photographic, heightmap and curvature. Each column shows derivatives of those main sources in the form of edge enhancements. Each feature serves as input for a 2D landmarker. The face used for this illustration is that of author MAdJ.

6.2.4 Synthetic symmetrical atlas creation

In order to construct our synthetic symmetrical facial surface model that will serve as our atlas, we take the average of the facial surface scans of 30 adults. To achieve this, we first automatically locate the 21 landmarks on each face using the aforementioned automatic landmarking algorithm. We then reshape all face models as close together as possible by applying Procrustes superimposition based on the 21 landmarks. This way, each individual face is reshaped to the mean 21 coordinates of all individuals.

After the superimposition step, 3D thin plate spline fitting aligns the 21 landmarks exactly by moving corresponding points into identical positions. Intermediate points are moved according to a bending energy and follow the movement of the anatomical landmarks thereby emulating a process of 'bending' one scan into the other. The resulting point cloud is then copied and mirrored horizontally, creating a symmetrical point cloud. This extremely dense point cloud is then thinned out by reducing the amount of points based on average local point density. The atlas model is then constructed using triangulation to add edges (Figure 6.5a). The final 3D synthetic face is created via a reverse map projection back to 3D (Figure 6.5b).

6.2.5 Region of interest

The atlas also assists in finding the region of interest, illustrated in Figure 6.6. Many 3D facial surface models include noise such as shoulders or hair styles that may have a detrimental effect on symmetry analysis. After fitting the atlas to each target face based on the 21 landmarks, a boundary is determined that serves as a cut-off to remove outlying data. This process is illustrated in Figure 6.7.

Figure 6.4: **An illustration of the set of 21 landmarks** displayed on a map projection of the face of author MAdJ.

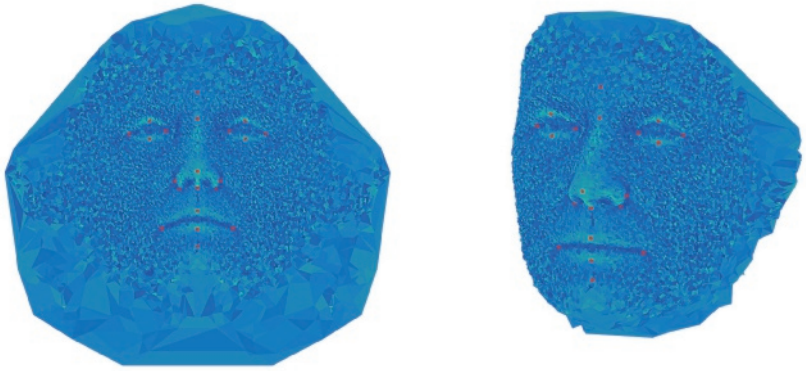
6.2.6 Symmetry registration

Once the region of interest is selected, these points are map projected to 2D together with the 21 landmarks that were registered earlier. The 2D face atlas is then aligned to the target face, again, using Procrustes superimposition based on the 21 landmarks. The 21 atlas landmarks are spline fitted to the 21 target landmarks, matching them exactly. The transformation induced by this spline fit is used to map the remaining landmarks of the atlas to the target's shape. As a result, the symmetrical atlas is bent into the shape of the target face. This provides a registration by virtue of the atlas annotation of the left and right symmetry of the face as illustrated in Figure 6.8.

6.3 Experiments

6.3.1 Data set

The dataset used to asses the performance of this algorithm is a random selection of 40 non-twin subjects from the *TwinsUK* cohort.



(a) Map projected state (b) 3D

Figure 6.5: **Average, symmetrical face atlas** in map projected state (a) and in 3D (b) with 21 reference landmarks in red.

The TwinsUK cohort consists of volunteers drawn from the general British population and is of of full European descent. The volunteers were unaware of any 3D studies scientific interests at the time of enrollment and gave fully informed consent under a protocol reviewed by the St. Thomas' Hospital Local Research Ethics Committee. Reference: PMID 23088889.

The *TwinsUK* dataset has models with ca. 1.5×10^5 points and textures with resolution of ca. 2000x1000 pixels. The data set was acquired with *3dMDface* photogrammetric systems.[1]

Data Availability

Due to privacy restrictions, raw data (facial 3D scans) cannot be made available for download. Subject to evaluation of a research proposal, the TwinsUK data-set is made available by co-authors Pirro Hysi and Tim Spector.

6.3.2 Single point deformation

In order to test symmetry registration accuracy, we add controlled deformations to a target face, in this case the original atlas, perform the registration process and evaluate. We perform this deformation firstly displacing a single landmark and performing a thin-plate-spline based adjustment of the remaining landmarks. In separate experiments, a single landmark is displaced in 6 different directions and for 3 distances: left-right, up-down, and front-back, with distances of 5, 10 and 15 mm. In total, 18 deformations simulations were performed per point. Example single point deformations in the 6 directions are illustrated in Figure 6.9.

For the experiment, one half of the face is overlaid with a regular grid of 50 points that serve as test locations for the deformations for our experiment. This grid is illustrated in Figure 6.10. After each

Figure 6.6: **Illustration of general region of interest** in relation to total available data.

deformation simulation, the symmetry of the resulting face is registered using the atlas.

6.3.3 Sagittal plane deformation

Besides local sources of asymmetry, facial non-normality can also entail the whole face. An interesting case is sagittal plane asymmetry that is also the result of some syndromes such as craniofacial microsomia (CFM). In the following experiment, we test the accuracy of our registration by artificially introducing a sagittal bending of the atlas.

The simulation process is initialized with a set of several selected points on the sagittal line on the surface of the face. These include points from the set of 21 landmarks that are already on the sagittal plane together with 4 new bending assist points. To simulate a natural sagittal bending, these points are displaced on the horizontal plane in such a way that horizontal distance of these points to the original sagittal line increases exponentially towards the bottom. In total, 5 different bending values of increasing distance were used in the experiment. Example sagittal deformations are illustrated in Figure 6.11. After each sagittal deformation simulation, the symmetry of the resulting face is again registered using the atlas.

6.3.4 Accuracy estimation

In order to estimate registration accuracy, we compare the meshes of the deformed model (truth) with landmarks derived from the registration process (observation). We do this by calculating the Hausdorff distance (equation 6.1), an established method to estimate the distance between two 3D meshes that do not contain the same set of points. Due to the large amount of pairwise distances to be computed, the Hausdorff distance calculation is approximated by drawing a random sub-sample of 25.000 landmarks

(a) Target surface scan with 21 landmarks

(b) Atlas surface scan with 21 landmarks

(c) Atlas landmarks (cyan) fitted to original landmarks (blue)

(d) Fitting result with atlas

(e) Selection boundary created based on atlas

(f) Final ROI selection result

Figure 6.7: **ROI selection process.** The process is illustrated with the face of author MAdJ.

from the surface of the mesh.

(a) (b)

(c)

Figure 6.8: **Symmetry registration shown in different views.** Yellow points indicate right-hand side of the face, purple the left. The face used in this illustration is that of author MAdJ.

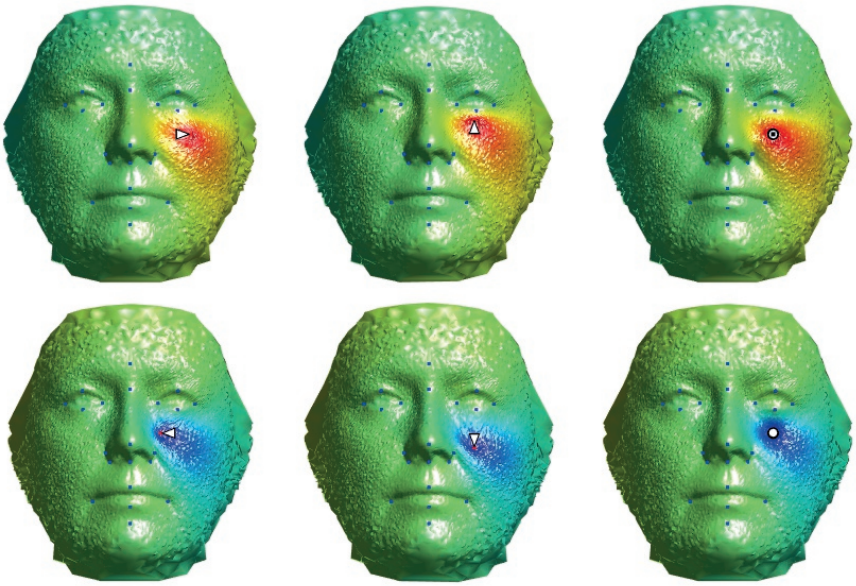


Figure 6.9: **Illustration of 6 single point deformations of the atlas face model.** The bending direction is denoted by the black arrow. A red color gradient illustrates a decrease distance in the direction of the arrow axis of points in relation to the original face model, blue color positive.

$$d_{\mathrm{H}}(X,Y) = \max\{ \sup_{x\in X} \inf_{y\in Y} d(x,y), \sup_{y\in Y} \inf_{x\in X} d(x,y) \}, \tag{6.1}$$

With $d(x,y)$ as distance function.

To counter any (unintended and unrelated) long-distance distortions as a result of 3D bending simulation method, a Gaussian-based weighting is applied to distances measured from the bending points. A density function with a standard deviation (*i.e.* Gaussian root mean square width) of 12.5 was used. This weighting is illustrated in Figure 6.12.

6.4 Results

6.4.1 Single point deformation results

The single point deformation results are illustrated in Figure 6.13. Generally speaking, the results show that bending points at the center of the face show the best registration accuracy ($<$ 0.1 mm). The lower left edge shows reduced accuracy for all deformation directions and distances ($>$ 0.1 mm), especially downwards at 15 mm. (0.4 mm). The upper edge of the face also shows, to lesser extent, reduced accuracy (0.1-0.2 mm), especially the outward displacement of 15 mm (0.3-0.4 mm). The upper left

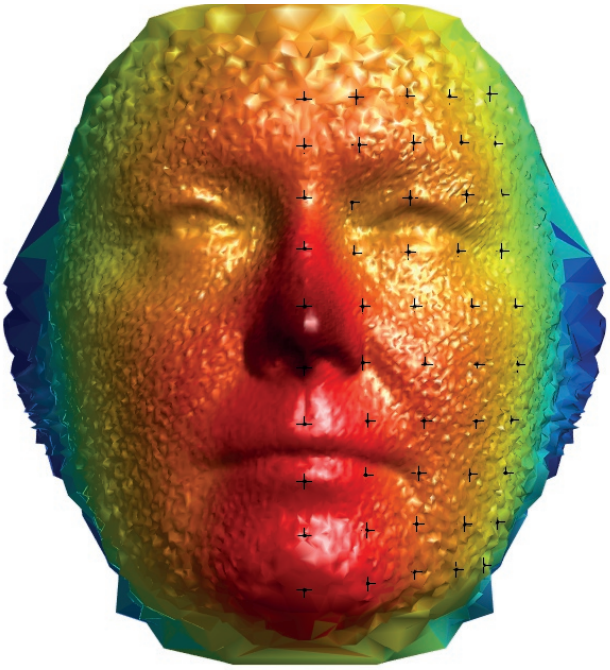


Figure 6.10: **Illustration of the grid used for the 50 locations for the single point deformation.**

corner of the face shows reduced accuracy all directions and distances (0.1-0.2). The upper left corner shows markedly less accuracy for the outwards direction, 15 mm (0.3-0.4 mm).

The sagittal plane deformation results are illustrated in Figure 6.14. When comparing the simulation and its registration in the first and second columns respectively and with the horizontal distance plotted on the third column, it can be concluded that the sagittal bending is best registered in proximity of the bending landmarks. This is shown by the fact that the border between left and right parts of the face closely traces the bending of the landmarks. With a maximum of 0.2 mm, the Hausdorff distance between simulation and registration shown in the rightmost column is small.

Sagittal plane deformation results

6.5 Discussion

In this paper, we investigated the accuracy of our automated facial symmetry registration algorithm with several simulation experiments. The registration algorithm relies on our previous automated landmarking algorithm that showed good accuracy for 21 landmarks (<2 mm). We performed simulation experiments in order to investigate the registration accuracy for single point deformation as well as global deformation by sagittal plane bending.

Results for the point deformation simulations displayed in Figure 6.13 show that registration accuracy is excellent in the regions near the registered 21 landmarks (< 0.1 mm) and still remain within 0.2-0.3

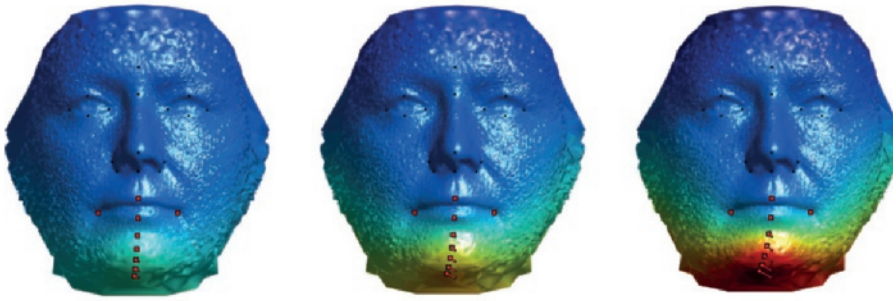
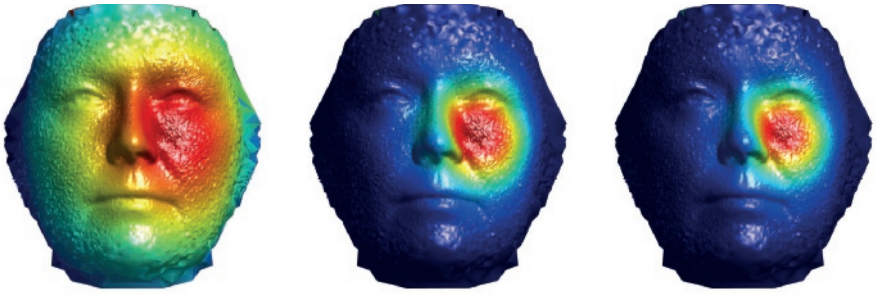


Figure 6.11: **Illustration of 3 iterations of increasing saggital plane deformations of the atlas face model.** The bending direction is denoted by the black arrows.



(a) Distance to deformation point (b) Gaussian filter (c) Gaussian filter applied

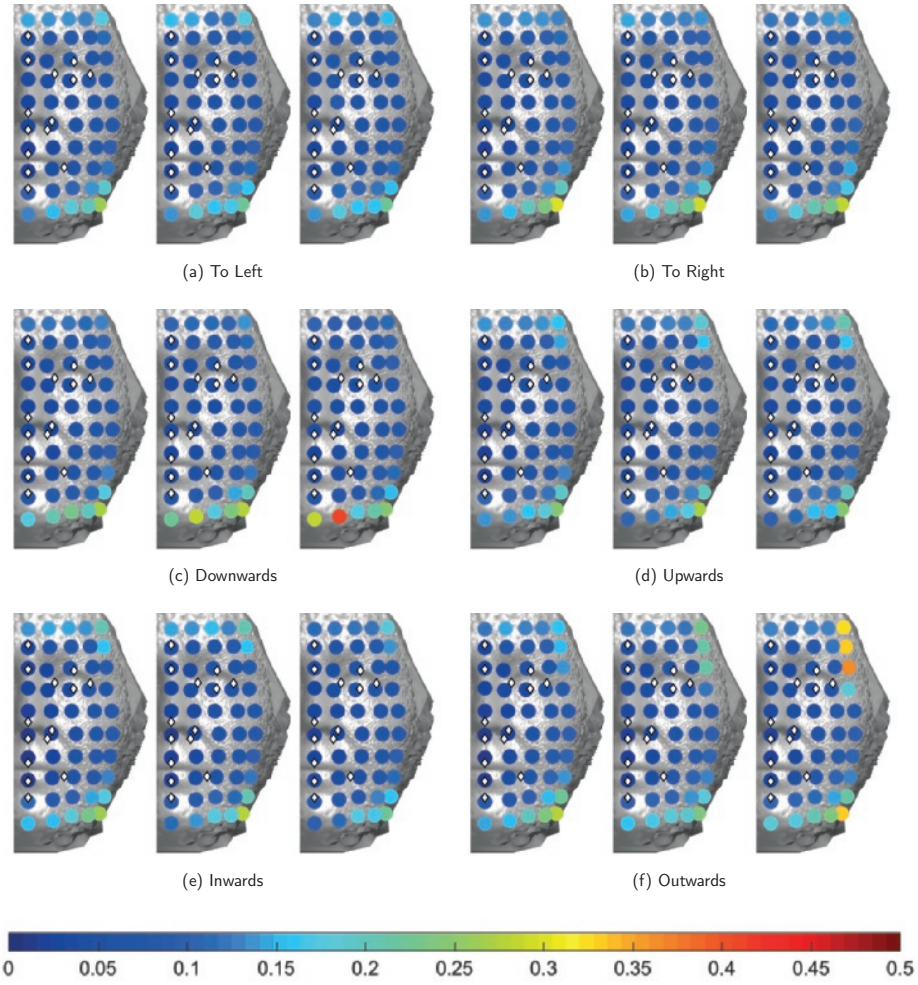
Figure 6.12: **Gaussian weighting at a single deformation point illustrated on the atlas face model.**

mm when inspecting registration accuracy in the outside regions of the face. Even the most extreme distances around the rim of the surface model are relatively small (ca. 0.3 mm) in comparison with the landmarking accuracy of < 2 mm.

Results for the sagittal deformation simulations show that surface registration distance was once more low (< 0.2 mm) and that sagittal bending can be accurately determined as long as enough landmarks are available. This setting is important for clinical applications.[2] We consider the fact that our proposed method strongly depends on the registration process as a major advantage. When improving landmarking accuracy, atlas based asymmetry estimation will automatically improve in accuracy as well. We used a registration method that is very flexible in the choice of landmarks and allows for the quick expansion of landmark sets. Other landmarking methods could of course also be used.

For our deformation experiments, we used the 3D surface corresponding to the atlas itself. Still, the landmark registration process was applied to this target image. Our results therefore disentangle symmetry registration from landmark registration by creating a situation where landmark registration is very accurate.

In our approach, symmetry registration accuracy is determined by the accuracy of the landmarking



(a) To Left (b) To Right

(c) Downwards (d) Upwards

(e) Inwards (f) Outwards

(g) Legend: mean distance in mm.

Figure 6.13: **Overview of the single point deformation results.** The results show Hausdorff distances between the original artificially bended model and its registration and are displayed in 3-item sets for displacement distances of 5, 10 and 15 mm (left to right). respectively. The landmark set is plotted in white diamonds.

algorithm. This can be a disadvantage, as other algorithms directly estimate asymmetry and provide a value for any mesh vertex. In these situations, correspondence between different samples is not established. Depending on the application this might be more efficient. However, in medical applications such as growth curves, correspondence between samples is required.

One way to improve accuracy of asymmetry estimation is to include more landmarks in the areas important for the application during the landmarking process. To achieve a more global assessment of the face, our current landmark set would have to be augmented by landmarks close to the rim of the face, such as the chin and the sides of the face. We plan to establish an optimal landmark set in future

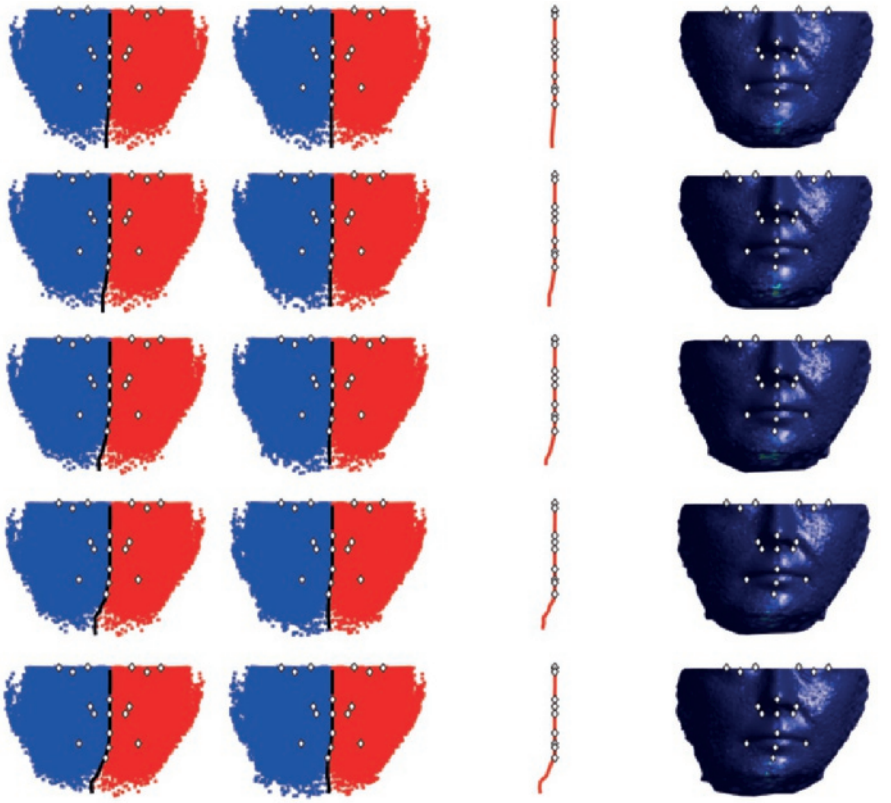


Figure 6.14: **Overview of the sagittal plane bending simulation results.** From top to bottom, each row represents a higher bending distance setting. The left column shows the original distribution of left and right side of the face after bending simulations. The second column shows the subsequent registration by our algorithm. The third column shows horizontal distance between the simulation original and its registration on the same scale. The final column shows Hausdorff distances between the surface of the original simulation model and its registration. The landmark set is plotted in white diamonds.

work. If more landmarks cannot be added, another solution would be to use principal components or regression to predict the bending direction further down the chin.

We did not compare our approach to competing methods. The main reason was the lack of available implementations. We provide the data sets used in the experiments, the atlas, and our own implementation in the hope of facilitating future comparison.

The ultimate judgment of facial symmetry lies in the eye of the human beholder and it is therefore of great interest to relate objective measures of asymmetry such as those recorded with our algorithm to human perception, even though this perception may vary from observer to observer.[7] Some experiments have been performed into what parts of the face contribute most to symmetry perception,[4] however, a focus on smaller details by concentrating on specific parts such as the nose or mouth would be most welcome. Another question is what degree of symmetry would be considered 'symmetrical enough', *e.g.*

what is the population mean for the perception of symmetrical normalcy? Such information would be helpful in *e.g.* deciding the medical necessity of surgery and setting surgery priorities. Addressing such question is beyond the scope of the current study but could be supported by our algorithm. Numeric asymmetry data derived from surface scans by our algorithm would have to be correlated with human rating data on the same scans under controlled circumstances. We plan to initiate such studies as they would have important clinical applications in predicting human perception of surgery outcomes.

In conclusion, we present an automated facial symmetry registration algorithm with guaranteed symmetrical correspondence that shows good accuracy of < 0.1 mm near (automatically located) landmarks and at most 0.2-0.4 mm for pseudo-landmarks with an applications in the measuring both local and global deformations of the face. Possible applications are population studies, growth curves, attractiveness perception studies, genetics, and medical applications such as surgery planning and surgery outcome prediction.

References

[1] *3dMD — 3D Imaging Systems and Software*. url: `http://www.3dmd.com/` (visited on 06/12/2014).

[2] Francesca Antonella Bianchi et al. "Soft, hard-tissues and pharyngeal airway volume changes following maxillomandibular transverse osteodistraction: Computed tomography and three-dimensional laser scanner evaluation". In: *Journal of Cranio-Maxillofacial Surgery* 45.1 (2017), pp. 47–55.

[3] Peter Claes et al. "Spatially-dense 3D facial asymmetry assessment in both typical and disordered growth". In: *Journal of anatomy* 219.4 (2011), pp. 444–455.

[4] Hyeon-Shik Hwang et al. "Three-dimensional soft tissue analysis for the evaluation of facial asymmetry in normal occlusion individuals". In: *The Korean Journal of Orthodontics* 42.2 (2012), pp. 56–63.

[5] Markus A Jong et al. "Ensemble landmarking of 3D facial surface scans". In: *Scientific reports* 8.1 (2018), p. 12.

[6] Mao Li et al. "Rapid automated landmarking for morphometric analysis of three-dimensional facial scans". In: *Journal of anatomy* 230.4 (2017), pp. 607–618.

[7] Joanna E Scheib, Steven W Gangestad, and Randy Thornhill. "Facial attractiveness, symmetry and cues of good genes". In: *Proceedings of the Royal Society of London B: Biological Sciences* 266.1431 (1999), pp. 1913–1917.

[8] L Spreeuwers. "Fast and Accurate 3D Face Recognition Using Registration to an Intrinsic Coordinate System and Fusion of Multiple Region". In: *Proc of Int Journal of Computer Vision* 93.3 (2011), pp. 389–414.

[9] Yi Sun and Lijun Yin. "Automatic pose estimation of 3d facial models". In: *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE. 2008, pp. 1–4.

[10] Helena O Taylor et al. "Quantitative facial asymmetry: using three-dimensional photogrammetry to measure baseline facial surface symmetry". In: *Journal of Craniofacial Surgery* 25.1 (2014), pp. 124–128.

[11] Randy Thornhill and Anders Pape Møller. "Developmental stability, disease and medicine". In: *Biological Reviews* 72.4 (1997), pp. 497–548.

7

General discussion

7.1 Introduction

The aim of this PhD project was to design, implement and apply automated landmarking methodology for use on 3D facial surface data. With an eye on intended applications such as GWAS analysis, surgery planning, and population studies, landmarking accuracy was of paramount importance alongside with flexibility for handling a broad range of heterogeneous data sets, and in combination with minimal training effort. Finally, downstream applications such as symmetry estimation should be made possible by this work.

We will now discuss each of the papers that were completed during this PhD, starting with the initial landmarking algorithm, followed by the improvements made possible by ensembles, an application to clinical data, an algorithm refocused on 3D human skull data, and finally a downstream application in the shape of a symmetry registration algorithm.

Next, we will deliberate on the methodological considerations of our algorithms, followed by a discussion of future work and an overview of new developments.

7.2 Main findings

7.2.1 3D landmarking with 2D Gabor Wavelets

In this work, we have successfully developed an automated approach for the landmarking of human facial 3D surface data. This algorithm applies a 3D to 2D map projection step to make use of an established 2D landmarking algorithm based on 2D Gabor wavelets.

Our results indicated that for some landmarks excellent accuracy was achieved with absolute differences between automatically and manually placed landmarks below 2 mm for most landmarks. Based on twin-related heritability estimates, our results showed that for the highest heritabilities, our algorithm could not be outperformed by a human rater. However, for certain distances, manual performance was much better. In summary, the distances of landmark results to true positions in the training data generally lie between 1-2 mm, which is competitive with other methods.

When evaluating our algorithm and comparing it with a state of the art implementation of an establish out-of-the-box landmarking algorithm (based on Active Shape Models (ASMs)), our algorithm showed superior results with regard to accuracy and ease of use. The latter lies in the fact that our implementation requires far less training data (our 30 vs. their 240 or even 3000+ samples) that allows for great flexibility when (re-)training a data set.

While, as said, general performance was very good, some landmarks performed better than others. Increasing this performance was the goal of our next iterations of the landmarking algorithm.

One possible cause of this performance difference is that for some landmarks the stack of 2D informative features that serve as input for our algorithm (*i.e.* photographic information, heightmap and edge enhanced transformations thereof) was insufficient. As our algorithm is quite modular and supports adding features without issue, it would be a relative cheap solution to add both new features (*e.g.* surface curvature information) and new transformations.

Another possible approach was inspired by the fact that for certain landmarks, certain features attributed positively to landmarking performance while others did less so. Here, we concluded that some kind of automatic feature selection that selects the best performing features for each landmark would improve landmarking performance. We therefore began to look into possibilities for such automated model selection methods.

A further challenge was the performance for landmarks with little underlying structural information or lack of edges in which the two before-mentioned suggestions would not be able to improve performance, *e.g.* located on smooth surfaces such as the forehead and chin. Inspired by other facial (2D) landmarking algorithms such as ASM, we decided to investigate the use of principal components (PCs). Such landmarks would benefit from efficiently limiting the search space for landmarks using PC predictions based on the training set.

One of the unique selling points of our algorithm is the non-heuristic nature of our approach, especially in comparison to other existing methods. This property should allow applications to other 3D surface than faces. The idea for 3D skull landmarking took root at this time. The flexibility of our landmarking and the fact that 3D skull landmarking is a far less researched topic with only a few

published methods made the realization of a 3D skull landmarking highly interesting.

In conclusion, our proposed method for automatic landmarking of facial 3D surface data required little investment in the training phase (ca. 30 training samples of 1 minute each for 21 landmarks) to automatically landmark 3D faces in a single iteration. Our experiments showed good performance (1-2 mm distance for most landmarks), in a dataset of faces of different quality, gender and ethnicity. The algorithm can be easily and quickly re-trained when a different set of landmarks is required. These properties are important for the landmarking of large medical 3D (facial) data sets.

There was also room for improvement which motivated us to start looking into expanding the feature set, automatic model selection and PCs for landmarks on smooth surfaces.

7.2.2 Ensembles

In our next paper, we describe the realisation of an improved landmarking algorithm that uses model selection based on *ensembles*. An ensemble method is a machine learning technique that combines the results of multiple learning algorithms into one to arrive at an even better result [15]. Such ensemble methods allow for a highly flexible range of input algorithms, as long as the intermediate results (*i.e.* the landmark coordinates) are of the same format.

In our work, we used ensembles to combine results from different landmarking algorithms into one in order to improve algorithm performance and to add automated feature selection mentioned in the previous section.

We were also able to implement a PC-guided search to tackle the problem of landmarks that have little structural information. This improved previously challenging landmarks in our set located on smooth surfaces such as the chin and forehead.

Selection in the ensemble formation ensures that for a given landmark only useful information is gathered from base landmarkers that in turn make use of specific features. We performed extensive experiments with two ensemble methods, *bagging* (model averaging) and *stacking* (regression).

Two experiment were designed to investigate the effectiveness of bagging and stacking for our needs. A major challenge for bagging was that it required 60,000 iterations of our algorithm. Normally, our algorithm was run on a desktop PC, but for such large scale computations, the LUMC computer cluster was required, including special code for job management. Ultimately, bagging did not improve results. Stacking, however, showed marked improvements by applying a linear regression on the coordinates.

Overall, this implementation achieved an average accuracy of 1.7mm, a 22% improvement over our previous algorithm.

Furthermore, in comparison with another automated landmarking method with a comparable landmark set, our improved ensemble algorithm shows better overall performance (2.6mm vs. 1.7mm). This positive comparison also holds when inspecting their best-performing individual landmarks: landmark 7 (tip of the nose) (1.6mm vs. 1.4mm), and landmark 13 (1.6mm vs. 1.5mm).

Deep networks can also be seen as ensembles as they also result in a classifier method based on different inputs. The main difference is that they can operate unsupervised and that their 'ensembles' are hidden in a black box design. The main limitation is the high number of training samples in comparison with our method, which we will address later. Nevertheless, deep networks are highly

relevant in computer vision today and in some way using deep network methods for our needs seemed worth investigating at this point.

At this time, we felt that our ensemble algorithm performed well enough to apply it on the ERGO data set for GWAS facial shape research. First, however, a field test was required. Therefore, we applied our automated algorithm on set of ca. 1500 TwinsUK subjects. Out of these 1500, 200 were manually labeled, which gave us a ground truth for direct accuracy comparisons. We were also able to use twin heritability information to perform heritability analyses and gain insight on landmark accuracy for the complete set of 1500. When the TwinsUK experiment confirmed that the algorithm performed just as well on new data as it had on our 30-40 item test in our papers, we proceeded with the ERGO set. Due to the large amount of 3D images (4430), and each face taking up to 15 minutes to complete (pre-processing + landmarking), it was decided to split the work over several desktop computers that ran the algorithm simultaneously for several days. For a small percentage of faces (4%), initial alignment had to be performed manually. After a visual inspection of the complete labeled set and removing obvious mislabeled results as well as invalid 3D images, we were left with 3838 labeled individuals. The results of this GWAS have been submitted the time of writing (2019).

7.2.3 A clinical application

Our third paper describes an application of our algorithm on clinical 3D facial data. In this combined study, sets of pre- and post-surgery images were quantified through 3D landmarking and compared. The investigated surgical procedures were mandibular mid-line distraction (MMD) and surgically assisted rapid maxillary expansion (SARME). This was performed for MMD (20 patients) and BiMEx, a combination of SARME and MMD (12 patients). Significant distance differences were found in line with surgery expectations. This practical application illustrates the potential of our algorithm research with small datasets.

Besides quantifying the effect of specific surgeries, the use of automated landmarking can give the clinical researcher a tool to quantify certain facial syndromes or study growth curves in a young population or investigate symmetry, of which the latter example will be discussed in the coming section.

Although evaluative research on such orthognathic surgery in 3D facial surface scans has been performed before, it it either used surface-based comparisons,[19] and when it was landmark-based, the landmarks were placed manually.[5,10] As such, our automated approach for research on orthognathic surgery is a novelty at the moment of publication.

7.2.4 Skull landmarking

As described in our fourth paper, we successfully applied our 3D ensemble landmarking algorithm for the automatic landmarking of CT scans of human skulls. The similarities of the data with facial data, the flexibility of our algorithm but also the near absence of automated skull landmarking methods in the scientific field motivated us to investigate this direction.

Firstly, leave-one-out experiments were performed with a training set of 30-40 conic CT scans of human skulls. These CT scans were first converted into surface models comparable to the 3D facial

scans. When evaluating accuracy with respect to multiple raters, 9 landmarks (out of an experimental set of 33) perform with absolute average discrepancy between automatically placed and trained landmark of below 2 mm. These results indicate that the algorithm is able to mimic a rater just as well as a rater can mimic him- or herself. An inter-rater comparison also shows nine well performing landmarks between them. Also, our algorithm reduced standard deviations, which further suggests better performance of our algorithm over manual landmarking.

Secondly, we have demonstrated in a practical experiment that skull superimposition, an important clinical application, is possible with our algorithm, although a manual step is required due to the limitations of the commercial software that performed the actual superimposition.

To summarize, our algorithm shows robust performance for 9 out of 33 landmarks with an accuracy of <2 mm. This is comparable with the best performance in our algorithm for human faces, and markedly better than another automated algorithm for human skulls (3.6–5.6 mm).

When further comparing our method with alternative, voxel-based approaches, many current studies still rely on manual placement. When considering manual labeling costs, this would 1) require the time of a trained expert, and 2) would take at least an estimated 1 minute per landmark, meaning that costs would sky-rocket for 100+-size data sets. In the case of automated voxel-based landmarking methods, some atlas-based methods exist,[2,16] however, these are limited by the low number of supported landmarks and reduced accuracy in comparison with our method. Alternatively, the development of a new voxel-based method would mean investing in a whole new project.

In conclusion, these results show that applying our algorithm to skull landmarking is a direction that is certainly worth investigating further.

7.2.5 Symmetry

The registration of facial symmetry is an important clinical application that is of great interest in relation to facial surgery and the study of surgery and syndrome development in relation to growth curves. We proposed to automate this registration by basing it upon our ensemble landmarking algorithm. This way, we were able to efficiently make use of a framework that was already available and that was able to give us the required accuracy and flexibility to create our symmetry registration algorithm.

Based on several simulation experiments, we conclude that the accuracy of our registrations on single point deformations, *i.e.* the surface distance between the original surface and its registration, is excellent in the regions near the deformation point (< .1mm) and still remain within .2-.3mm in other areas. Even extreme distances at the edges of the face (>.3mm) are relatively small in comparison with the landmarking accuracy of <2mm.

Our experiments with sagittal bending also shows surface distances <.2mm. Furthermore, we conclude that left-right side registration accuracy was excellent in areas that had sufficient supporting landmarks present.

In conclusion, we present an automated facial symmetry registration algorithm with guaranteed symmetrical correspondence with good accuracy of <.1mm near registration landmarks and correct recording of sagittal plane deformations.

7.3 Methodological considerations

In this section we discuss the methodological limitations and considerations of our landmarking algorithm and possible alternative approaches and solutions.

7.3.1 Inherent data type limitations

Photogrammetrical surface date

A first limitation to consider is that the quality of the input data dictates the quality of the 3D landmarks. To reiterate, the 3D facial surface data sets used during this thesis consists of triangulated 3D points and their connecting edges. On the surface created by these connected these points, a color photograph is projected.

Both of the data categories, the 3D coordinates and the photographic information, have their limitations that often go hand in hand due to their shared photogrammetrical origin. An example of low-quality issues are low detail, unfortunately enough often in highly informative areas such as eye corners, the edges of eyelids and mouth corners. Such low quality manifested itself in blurry textures and 3D artifacts.

These problems are often caused when multiple photographs taken from different angles have to be merged into the same point on the surface. This is especially prominent in concave areas where photographs from different angles meet, such as the eye corners. Such meeting points may result in jagged tears and blurriness. Other difficult areas are transparent areas, *e.g.* in the eye, or semi-transparent features such as eyelashes where conflicting 3D information from multiple cameras has to be merged. These situations may lead to 3D artifacts, such as spikes. Advances in technology such as high(er) resolution photography or laser scanning may reduce this problem in the future.

Ideally, the dataset would consist of surface scans that have an error $<.2$ mm for difficult 3D locations, together with individual high-resolution (ca. 4000x4000 pixels or larger) photographic maps per camera. To reduce 3D artifacts, shape information from CT, MRI, laser scanning or other high-resolution techniques could be merged with the photogrammetrical data. As for the region of interest, ideally, the 3D images should include also the back of the head and neck to support research into *e.g.* head circumference or the shape of the ear.

CT data

An important issue specific to CT data is the presence of artifacts. In our data, the largest contributor to this issue is scatter caused by metal dental restorations resulting in blobs and spikes. The presence of these artifacts is the main reason why no landmarks were tested in those areas. Another factor that limits accuracy is the general quality of the image, where low-resolution CT images tend to result in larger structural cavities in the image *e.g.* in the zygomatico-maxillary suture area (the cheek bone). Finally, a specific limitation in cone beam data is the horizon of detail that causes fading detail towards the lateral side of the skull. In our application, this effect was not relevant as most of our landmarks are located close enough to the medial center of the skull.

Using the recently developed high-resolution multislice computed tomography (MSCT) could improve the subsequent landmark quality. Also, special CT techniques exist that alleviate the problem of artifacts called metal artifact reduction (MAR) [14].

7.3.2 Initial registration

Our algorithm depends on properly aligned faces and skulls: better initial alignment of unprocessed 3D facial data, or registration, improves map projection and subsequent landmarking performance. To be specific, properly aligned surface models will stabilize the landmark search grids. Several methods have been applied to find the optimal registration, however, non-optimal alignment and an over-reliance on heuristics still remains a drawback of our method. For example, when processing data from ERGO, our alignment algorithm was able to align 96%, the remaining faces had to be aligned manually.

7.3.3 Map projections

The map projection of faces from 3D to 2D is a characteristic of our algorithm. During this projection, some landmarks, *e.g.* the base of the nose between the nostrils, may become occluded, especially for certain (aquiline) noses in the population. The extent of this issue, however, appears to be limited only to specific cases.

7.3.4 Running time optimization

As the development process of our landmarking algorithm may be characterized as creating increasingly functional prototypical pipeline, optimization was never a goal nor a necessity. This has lead to sub-optimal computational processes resulting in long (pre-)processing times. Although the algorithm could be sped up by using high-end hardware or parallelization, investments in software optimization would be a more effective solution. Examples could be to rewrite code to use more efficient methods or to perform the feature creation on the video card memory directly. Another option would be to migrate the algorithm to more efficient programming language altogether.

7.3.5 Ethical considerations

The landmarks that are generated by our algorithm are a potential privacy risk. Even though 3D representations in biometrics are currently rare, this will most likely change in the future, *e.g.* the current iPhone X already uses 3D facial verification based on infra-red point projections [6] and some prototype models support time of flight 3D imaging [22]. More importantly, it is trivial to generate 2D representations from 3D landmarks. 2D landmark representations are much more common and can be easily obtained from photographs and compared, posing a serious privacy breach.

Data leaks of such data might also bear a risk similar to genetic data as it is impossible to start over with a 'fresh copy'. This is an argument against the indiscriminate collection of facial data.

Furthermore, in the training phase, a set of 30-40 subjects are used. Besides the 3D training landmarks, the photographical and other structural 2D data are convoluted with Gabor wavelets. These

convolutions are stored and subsequently used in comparisons with to the test set in the landmarking application phase. The training set, however, can be reconstructed from the stored convolutions, and in combination with the 3D landmarks a 3D reconstruction of the whole face may be possible. It is therefore important to treat the training set as highly confidential and to apply sound security, *e.g.* encryption, to this data and to store it as 'cold data' (*i.e.* on disconnected hard drives).

A potential but important issue that also relates to the training phase of our algorithm is that a bias towards the ethnicity of the selected training set population may emerge. Although skin tone is only present in a small subset of features based on photographic information and should not be consequential, ethnic variations in 3D shape of *e.g.* the nose, eye and mouth may play a somewhat larger role as these variations permeate into all 2D features. Even though we are not aware of such biases in our current results, they are expected to be present in some form. A consequence of biased training sets could be biased studies, *e.g.* under-represented elements of the face would receive less accurate landmarks and in turn negatively influence scientific conclusions drawn from them.

Finally, clinical, forensic and security applications of 3D landmarks may assist a democratic nation in keeping its population healthy and secure. However, during the development of such algorithms, one must always be aware of its potential in non-democratic nations where its usage may contradict with our values on privacy and non-discrimination and our local laws in general.

7.4 Future work and new developments

Just as after every other research project, there is always room for improvement. For our landmarking algorithm and algorithms based thereon, we will take a look at possible future directions and discuss new development taking place at the time of writing.

First of all, the initial registration phase in which the faces are aligned before they are landmarked remains a research question in its own right. As such, it has received less attention in this project as the focus was on landmarking. Even though our results show that our current algorithm is robust enough to deal with non-perfectly aligned data at later stages of the pipeline. Even with a success rate of 96%), there is still some room to further optimize the initial registration methods and to make this phase less heuristic as well. We can also opt for an external alignment method, or to combine it with our own method in a kind of ensemble. A candidate registration method using deep learning is focused on 2D but uses a 3D reconstruction model [3].

Secondly, as we have demonstrated, our algorithm is very modular, expressed both the ability to expand the stack of 2D features and by being able to combine completely different methods into an ensemble.

Even though the current feature set is informative enough for most landmarks, it is still attractive to simply expand this stack further. For example, additional edge detection filters can be added. One might also focus on specific colors of the photographic feature for *e.g.* detecting the edge of the lip by making use the contrast of its red color with normal skin. A more advanced option to record additional 3D information is to record 2D features from different point of views of the same 3D landmark, or changing the lighting conditions (*e.g.* direction, strength) so that shadows being cast improve contrast

for some features.

After our second iteration, we are able to include the results from external landmarking algorithms into an ensemble. Examples of suitable methods include pre-existing and pre-trained 2D active shape model implementations such as STASM [12], 3D atlas based methods [11], 3D dense correspondence registration methods [8, 9] and deep learning implementations in 2D [17] and 3D [23]. However, we will most likely be limited to overlapping landmark sets due to the aforementioned differences in training set size. Of course, there are other considerations as well, such as accuracy and heuristics. While our algorithm requires a set of 40 manually labeled training samples, atlas based methods 50, 3D dense correspondence methods require 200, active shape models require 3000+ and deep learning even requires 25,000+(!). Training costs therefore may rise considerably when deciding to include non-pre-existing landmarks from such methods, which would work against the advantages of our own algorithm. Ultimately, from all of these landmarking algorithm candidates, the optimal combination for each landmark would selected via stacking.

Specific plans for the skull landmarking algorithm are further experimentation with this type of data to bring the results for all 33 landmarks to the level of the 9 well-performing ones. As suggested before, this may be achieved by making sure that a large enough training set (30-40) is available for all landmarks. The suggestion of creating 2D features from different perspectives might be more effective here than for facial data due to the more extreme convex and concave surface of the skull, leading to more frequent occlusions of landmarks. Although not many automated skull landmarking methods exist, incorporating other methods, *e.g.* a recent 2D active shape model method [13], into the ensemble can improve its results for skulls comparably likewise as for facial data.

The positive results with clinical data in our pre- vs. post-surgery comparison, encourages us to further develop our algorithm to support such research. In its current state, our algorithm includes a point-and-click 3D facial labeling tool created in Matlab which allows clinicians to train landmarks and to inspect and correct results. However, the steps required to feed the raw data into the algorithm, starting the pre-processing and landmarking processes and retrieving the subsequent results are still initiated manually and were performed by the author for this study. Ideally, streamlining the complete process into a (web-based) graphical user interface is something that must still be realized.

As reported in our main findings, using automated landmarking for soft tissue analysis in orthognathic surgery evaluation is a novelty. However, a combination of our automated soft tissue surface model algorithm with our hard tissue skull algorithm would be highly interesting for these purposes. Effectively, this could automate what is now a highly labor-intensive area of research: investigating the changes caused by surgery in both soft and hard tissue.[5]

The symmetry registration algorithm is limited by its landmark set. Expansion of important landmarks, *e.g.* on the sagittal plane will achieve best results for left-right side symmetry registration. Furthermore, performing experiments that compare the symmetry results of our algorithm with the perception of human raters are of great interest, as well as investigating what parts of the face contribute to the symmetry perception.[19] Moreover, such information can assist in defining a way to summarize the symmetry results in a human-readable form.

Deep learning techniques are an interesting alternative to our landmarking methods as these can potentially work on raw data directly and do not require specification of features.[24] However, as stated

before, in contrast with our landmarking algorithm, a up to a thousand-fold are required for deep learning.[3,24] We could still, however, use the strengths of deep learning without leaving behind our algorithm's small training set advantage by limiting the number of features, as well as the number of layers in the network. Also, we could re-purpose parts of deep learning algorithms (*i.e.* transfer learning[4]) to make our method more flexible and generic whilst retaining the advantage of our small training set. Investigating the incorporation of such deep learning elements into our algorithm could be highly advantageous.

One example of such deep learning techniques is GoogleNet[21] that has been used in *e.g.* face recognition.[20] GoogleNet is capable of dealing with highly heterogeneous data, one of the great advantages of deep neural networks. Work on applying deep learning specifically to 3D voxels has also been reported [25].

Another interesting recent development in 3D anatomical landmarking is the use of artifical agents that combine behavioral learning to optimize a search strategy with deep learning-based feature extraction for object appearance detection [1, 7]. In short, this allows the artificial agent to "navigate" a path in 3D towards the point of interest.

Finally, an emerging problem in artificial intelligence and computer vision is the problem of bias [18]. As we discussed in methodological considerations, our algorithms have the potential to propagate bias that *e.g.* is introduced during the selection of the training set. It is of great importance to investigate such biases and to find ways to guarantee correct results indiscriminately of *e.g.* subject ethnicity, age or gender.

References

[1] Amir Alansary et al. "Evaluating Reinforcement Learning Agents for Anatomical Landmark Detection". In: (2018).

[2] Stefan Boehringer et al. "Genetic determination of human facial morphology: links between cleft-lips and normal variation". In: *European Journal of Human Genetics* 19.11 (2011), pp. 1192–1197.

[3] Adrian Bulat and Georgios Tzimiropoulos. "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)". In: *International Conference on Computer Vision*. Vol. 1. 6. 2017, p. 8.

[4] Philippe Burlina et al. "Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis". In: *Computers in Biology and Medicine* 82 (2017), pp. 80–86.

[5] Yu-Jen Chang et al. "Soft Tissue Changes Measured With Three-Dimensional Software Provides New Insights for Surgical Predictions". In: *Journal of Oral and Maxillofacial Surgery* 75.10 (2017), pp. 2191–2201.

[6] Steven Ecott. *iPhone X: The end of privacy?* 2017. url: `https://mastersofmedia.hum.uva.nl/blog/2017/09/23/iphone-x-the-end-of-privacy/` (visited on 06/06/2018).

[7] Florin C Ghesu et al. "Towards Intelligent Robust Detection of Anatomical Structures in Incomplete Volumetric Data". In: *Medical Image Analysis* (2018).

[8] Syed Zulqarnain Gilani, Faisal Shafait, and Ajmal Mian. "Shape-based automatic detection of a large number of 3D facial landmarks". In: *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE. 2015, pp. 4639–4648.

[9] Carl Martin Grewe and Stefan Zachow. "Fully automated and highly accurate dense correspondence for facial surfaces". In: *European Conference on Computer Vision*. Springer. 2016, pp. 552–568.

[10] Mohammad Y Hajeer, Ashraf F Ayoub, and Declan T Millett. "Three-dimensional assessment of facial soft-tissue asymmetry before and after orthognathic surgery". In: *British Journal of Oral and Maxillofacial Surgery* 42.5 (2004), pp. 396–404.

[11] Mao Li et al. "Rapid automated landmarking for morphometric analysis of three-dimensional facial scans". In: *Journal of anatomy* 230.4 (2017), pp. 607–618.

[12] Stephen Milborrow and Fred Nicolls. "Active shape models with SIFT descriptors and MARS". In: *VISAPP* 1.2 (2014), p. 5. url: `http://www.dip.ee.uct.ac.za/~nicolls/publish/sm14-visapp.pdf` (visited on 06/05/2015).

[13] Jesús Montúfar, Marcelo Romero, and Rogelio J Scougall-Vilchis. "Automatic 3-dimensional cephalometric landmarking based on active shape models in related projections". In: *American Journal of Orthodontics and Dentofacial Orthopedics* 153.3 (2018), pp. 449–458.

[14] Andre Mouton et al. "An experimental survey of metal artefact reduction in computed tomography". In: *Journal of X-ray Science and Technology* 21.2 (2013), pp. 193–226.

[15] David Opitz and Richard Maclin. "Popular ensemble methods: An empirical study". In: *Journal of artificial intelligence research* 11 (1999), pp. 169–198.

[16] Roshan N Rajapakse et al. "Automated Extraction of Cranial Landmarks from Computed Tomography Data using a Combined Method of Knowledge and Pattern Based Approaches". In: *Applied Medical Informatics* 38.1 (2016), p. 1.

[17] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[18] Hope Reese. *Bias in machine learning, and how to stop it*. 2016. url: `https://www.techrepublic.com/article/bias-in-machine-learning-and-how-to-stop-it/` (visited on 06/06/2018).

[19] Ann-Sophie Storms et al. "Three-dimensional aesthetic assessment of class II patients before and after orthognathic surgery and its association with quantitative surgical changes". In: *International journal of oral and maxillofacial surgery* 46.12 (2017), pp. 1664–1671.

[20] Yi Sun et al. "Deepid3: Face recognition with very deep neural networks". In: *arXiv preprint arXiv:1502.00873* (2015).

[21] Christian Szegedy et al. "Going deeper with convolutions". In: Cvpr. 2015.

[22] Vivo. *Vivo Showcases Pioneering TOF 3D Sensing Technology at MWC Shanghai 2018*. 2018. url: http://www.vivo.com/en/about-vivo/news/vivo-showcases-pioneering-tof-3d-sensing-technology-mwc-shanghai-2018 (visited on 06/29/2018).

[23] Amir Zadeh et al. "Convolutional experts constrained local model for 3d facial landmark detection". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. Vol. 7. 2017.

[24] Zhanpeng Zhang et al. "Learning deep representation for face alignment with auxiliary attributes". In: *IEEE transactions on pattern analysis and machine intelligence* 38.5 (2016), pp. 918–930.

[25] Yefeng Zheng et al. "Robust Landmark Detection in Volumetric Data with Efficient 3D Deep Learning". In: *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*. Ed. by Le Lu et al. Cham: Springer International Publishing, 2017, pp. 49–61. isbn: 978-3-319-42999-1. doi: 10.1007/978-3-319-42999-1_4. url: https://doi.org/10.1007/978-3-319-42999-1_4.

8

Summary

The aim of this PhD was to design, implement and further improve upon an automated landmarking algorithm for 3D facial surface data based on 2D Gabor wavelets and to apply it to real world data.

The first chapter of this thesis introduces the reader to the problem at hand and starts with an introduction into computer vision through a thought experiment. As we discuss the challenges we meet, we move step by step closer towards the specification of the requirements of our landmarking algorithm.

The second chapter proposes the first implementation of our automatic landmarking algorithm. In it, we describe how 3D surface facial data is map projected to 2D. From these projections, we extract 2D information, or features, such as photographic and height maps and their derivatives. These 2D features serve as an input for a trainable 2D pattern matching landmarking algorithm based on Gabor wavelets. Once the landmarks are located, the coordinates are reverted back to 2D with no loss of accuracy. We are able to detect 21 meaningful anatomical landmarks. Our experimental results show that our algorithm requires minimal investment in the training phase (ca. 30 training samples taking 1 minute for each landmark) to automatically landmark 3D faces in a single iteration. As such, our algorithm can be easily and quickly re-trained when a different set of landmarks is required. Leave-one-out experiments show good performance (1-2 mm distance to manually placed labels for most landmarks) over faces of difference quality, gender and ethnic background. An additional twin heritability experiment also showed good results. In conclusion, our algorithm meets the requirements of having meaningful landmarks, good accuracy and flexibility that are important for the landmarking of large medical 3D (facial) data sets.

The third chapter presents an improved landmarking algorithm enhanced by three additions. Firstly, we included additional 2D features to give the algorithm more information to work with. Secondly,

we introduce the use of ensembles. During the formation of the ensemble, a selection takes place in which for a given landmark only useful information (*i.e.* features) is included with the goal of improving landmarking accuracy. Thirdly, we were able to improve landmarks with little underlying structural information by more accurately predicting the search area based on principal component analysis of the training landmarks. Based on leave-one-out experiments, all three additions proved successful. A twin heritability experiment once more again supported this conclusion. we achieved an average accuracy of 1.7mm, a 22% improvement over our first iteration and a better performance than competing algorithms. The ability to add additional features as well as other algorithms to our ensemble, makes our algorithm highly modular as well.

The fourth chapter describes a clinical application in the form of a pre- and post-surgery comparison. In a combined study on mandibular midline distraction (MMD) and surgically assisted rapid maxillary expansion (SARME) and BiMEx, a combination of both. Significant distance differences were found in line with surgery expectations. This practical application illustrates the potential of our landmarking algorithm in a clinical setting, even with with small datasets.

The fifth chapter introduces a landmarking algorithm adjusted specifically for human skulls, or rather 3D CT scans of skulls. This algorithm shows robust performance for 9 out of 33 experimental landmarks with an accuracy of $<$2mm. This is comparable with the best performance in a similar algorithm for human faces, and markedly better than another automated algorithm for human skulls. Even though the algorithm needs more work to support larger landmark sets, our results show that automatic skull superimposition, an important clinical application, is already feasible with the current iteration of our algorithm. This algorithm is an example of the flexibility of our algorithm as well.

The sixth chapter presents a facial symmetry registration algorithm that is supported by (our) landmarks. The main goal of this algorithm was to enable a registration with a guaranteed symmetrical correspondence between both sides of the face. We attempt to achieve this goal by using a synthetic symmetrical atlas (a *i.e.* face model) that has this correspondence and to subsequently 'bend' this atlas to the shape of the target face, supported by (automatically located) landmarks. Experiments show good accuracy of $<$ .1 mm distance between atlas and target face near supporting landmarks, and at most .2-.4 mm in areas further removed. We also paid particular attention to the registration of sagittal plane (*i.e.* vertical) deformation of the face in our experiments. Possible applications are population studies, attractiveness perception studies and genetics. Another likely application is the creation of symmetry growth curves over the years in normal vs. syndrome cohorts. Comparing such growth curves could help in determining the optimal moment of surgery in a young patient's life.

In conclusion, our landmarking algorithm has met all its predetermined goals by delivering meaningful, accurate and stable landmarks in combination with high flexibility in terms of short training times and modularity. Furthermore, the algorithm's flexibility is illustrated by a successful application on human skulls, a highly interesting research venue in its own right. Although large scale research based on landmarks that were delivered by our algorithm for facial genetics is nearing completion, we have already shown the algorithms potential with small data sets in clinical pre- versus post-surgery comparisons. A promising facial symmetry registration algorithm may open the door to additional future research.

9

Samenvatting

Het doel van deze PhD was het ontwerpen, implementeren en verder verbeteren van een algoritme gebaseerd op 2D Gabor wavelets voor het automatisch vinden van anatomische punten, of *landmarks*, op 3D-scans van het gezicht. De tweede doelstelling was het toepassen van dit algoritme op data uit het veld bij verschillende experimenten.

Het eerste hoofdstuk presenteert het op te lossen probleem aan de lezer en geeft een introductie van het vakgebied van computer visie aan de hand van een gedachte-experiment. Terwijl we stap voor stap de uitdagingen van dit veld behandelen, komen we langzaam maar zeker tot de specificatie van de vereisten van ons landmark-algoritme.

Het tweede hoofdstuk doet de eerste implementatie van ons automatische landmark-algoritme uit de doeken. Hierin wordt beschreven hoe de 3D-afbeeldingen geprojecteerd worden naar 2D. Uit deze projecties worden vervolgens verschillende vormen van informatie gehaald, namelijk het fotografisch oppervlak en de hoogteverschillen, en afgeleiden hiervan. Denk bij dit laatste aan filters die de weergave van randen in de afbeelding digitaal versterken. Deze 2D informatie dient als invoer van een trainbaar 2D-landmark-algoritme dat gebaseerd is op de werking van Gabor wavelets. Nadat de landmarks zijn gelokaliseerd, worden de 2D-coordinaten zonder verlies van precisie teruggeprojecteerd naar 3D. We zijn in staat om op deze manier 21 anatomische landmarks te lokaliseren. Onze resultaten laten zien dat ons algoritme een kleine investering in tijd nodig heeft (ca. 1 minuut voor elk van de 30 voorbeelden die moet worden gelabeld per landmark) om vervolgens automatisch 3D-gezichten te kunnen labelen zonder verdere handelingen. Het algoritme kan dus eenvoudig en snel opnieuw getraind worden wanneer een andere set landmarks nodig is. Uit *leave-one-out* experimenten laten goede resultaten zien (1-2 mm afstand tot handmatig labelen voor de meeste landmarks) voor afbeeldingen van verschillende kwaliteit,

geslacht en etniciteit. Een genetische tweelingstudie met betrekking tot erfelijkheid ondersteunt deze resultaten nog verder. Er kan dus worden geconcludeerd dat ons algoritme aan de volgende eisen voldoet: het geven van anatomisch betekenisvolle landmarks, uitstekende precisie en flexibiliteit. Dit zijn vereisten die belangrijk zijn voor het labelen van grote medische 3D data sets.

Het derde hoofdstuk omschrijft de volgende iteratie van ons algoritme aan de hand drie verbeteringen. De eerste is het toevoegen van extra bronnen van 2D-informatie om het algoritme meer informatie te geven om mee te werken. Als tweede introduceren we het gebruik van *ensembles* die automatisch alleen de 2D-informatie selecteert die nuttige informatie verschaft en de rest negeert. Hiermee wordt de precisie verhoogt. Ten derde is de prestatie voor landmarks op gladde huidoppervlaktes verbeterd door het gebruik van *principal components*. Hierdoor kan het algoritme betere voorspellingen doen op plaatsen die weinig structurele houvast bieden zoals bij landmarks op het kin of voorhoofd. Leave-one-out-experimenten hebben aangetoond dat al deze verbeteringen hebben geleid tot een hogere precisie, met een gemiddelde van 1.7mm; een verbetering van 22% boven onze eerste iteratie en bovendien beter dan concurrerende algoritmes. Verder zorgt de optie om eenvoudig meer 2D-informatie toe te voegen en de mogelijkheid om andere algoritmes aan ons ensemble mee te laten doen ervoor dat ons algoritme zeer modulair is.

Het vierde hoofdstuk omhelst een klinische toepassing in de vorm van een pre- versus post- operatievergelijking. In dit onderzoek worden de effecten van verschillende kaakoperaties in kaart gebracht door het meten en vergelijken van afstanden tussen de landmarks van vóór en na de operatie. Het betrof hier verschillende operaties van de boven- en onderkaak, namelijk *mandibular midline distraction* (MMD), *surgically assisted rapid maxillary expansion* (SARME) en BiMEx, een combinatie van de twee eerdergenoemde operaties. Er werden statistisch significante veranderingen gevonden die in lijn lagen met de verwachtingen. Deze praktische toepassing illustreert de mogelijkheden van ons algoritme in een klinische setting en met een kleine data set.

Het vijfde hoofdstuk introduceert een aangepast algoritme voor het vinden van landmarks op 3D-CT-scans van de menselijke schedel. Dit algoritme geeft robuuste prestaties voor 9 van de 33 geteste landmarks. Dit is vergelijkbaar met de beste prestaties van vergelijkbare algoritmes voor het menselijk gezicht en een stuk beter dan alternatieve algoritmes voor schedels. Hoewel er nog werk verzet moet worden voordat er grotere sets van landmarks ondersteund kunnen worden, kunnen er met de huidige resultaten al wel automatisch schedels over elkaar heen gelegd worden. Dit is een belangrijke klinische toepassing. Ook deze toepassing van het algoritme toont nogmaals de flexibiliteit van onze methode aan.

Het zesde hoofdstuk presenteert een registratie-algoritme van de symmetrie van het gezicht, ondersteund door onze landmarks. Het hoofddoel was het automatisch registreren van een gegarandeerde correspondentie tussen punten in de linker en rechterhelft van het gezicht. We hebben hier gekozen voor het maken van een kunstmatig gevormd atlas-gezicht waarin deze correspondentie al aanwezig is. Vervolgens wordt deze atlas in de vorm van het doel gebogen waarbij deze correspondentie aan het doel-gezicht wordt overgedragen. Uit experimenten met simulaties van kunstmatige verbuigingen van een gezicht blijkt dat deze registratie een nauwkeurigheid heeft van < 1 mm rond de ondersteunende landmarks en ten hoogste .2-.4 mm op grotere afstand. Ook is er een experiment uitgevoerd met gesimuleerde verbuigingen over de verticale as van het gehele gezicht. Mogelijke toepassingen van

deze registratiemethode zijn populatiestudies, experimenten met de menselijke perceptie van aantrekkelijkheid en genetische studies. Een andere voor de hand liggende toepassing is het vervaardigen van groeicurves die de ontwikkeling van de symmetrie van het gezicht over de jaren in kaart kunnen brengen. Met deze groeicurves kan bijvoorbeeld onderzoek gedaan worden naar het optimale tijdstip van het uitvoeren van een gezichtsoperatie van patiënten die in de groei zijn.

Concluderend kan worden gesteld dat ons algoritme alle genoemde doelen heeft gerealiseerd door het leveren van anatomisch betekenisvolle, nauwkeurige en stabiele landmarks in combinatie met een hoge mate van flexibiliteit in termen van korte trainingstijd en modulariteit. De flexibiliteit van het algoritme wordt verder geïllustreerd door een succesvolle toepassing op schedels, een onderzoeksgebied dat overigens op zichzelf ook zeer interessant is. Hoewel de resultaten van data die geleverd is voor grootschalig genetisch onderzoek nog gepubliceerd moeten worden, hebben we al wel aan kunnen tonen dat ons algoritme veel potentie heeft in kleinschalig klinisch pre-/post-operatieonderzoek. Een veelbelovend algoritme voor het registreren van symmetrie opent ten slotte de deur naar nog meer onderzoek.

Appendices

Author's Affiliations

Department of Oral & Maxillofacial Surgery, Special Dental Care, and Orthodontics, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands
Markus de Jong, Eppo Wolvius, Maarten Koudstaal, Atilla Gül, Jan Pieter de Gijt

Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands
Markus de Jong, Stefan Böhringer

Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands
Markus de Jong, Fan Liu, Manfred Kayser

Section of Evolutionary Biology, Department of Biology II, University of Munich LMU, Planegg-Martinsried, Germany
Andreas Wollstein

Department of Medical Physics, University College London Hospital, Londen, United Kingdom
Clifford Ruff, David Dunaway

Craniofacial Unit, Great Ormond Street Hospital for Sick Children, London, United Kingdom
Maarten J. Koudstaal, Pirro Hysi, Tim Spector

Department of Twin Research and Genetic Epidemiology, King's College London, United Kingdom
Pirro Hysi, Tim Spector

Department of Medical Informatics, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands
Wiro Niessen

Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands
Wiro Niessen

Publications

de Jong, Markus, Mariët Theune, and Dennis Hofs. "Politeness and alignment in dialogues with a virtual guide." In Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1, pp. 207-214. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

de Jong, Markus A., Andreas Wollstein, Clifford Ruff, David Dunaway, Pirro Hysi, Tim Spector, Fan Liu, Wiro Niessen, Maarten J. Koudstaal, Manfred Kayser, Eppo B. Wolvius, Stefan Böhringer. "An automatic 3d facial landmarking algorithm using 2d gabor wavelets." IEEE Transactions on Image Processing 25, no. 2 (2016): 580-588.

de Jong, Markus A., Pirro Hysi, Tim Spector, Wiro Niessen, Maarten J. Koudstaal, Eppo B. Wolvius, Manfred Kayser, and Stefan Böhringer. "Ensemble landmarking of 3D facial surface scans." Scientific reports 8, no. 1 (2018): 12.

de Jong, Markus A., Atilla Gül, Jan Pieter de Gijt, Maarten J. Koudstaal, Manfred Kayser, Eppo B. Wolvius, and Stefan Böhringer. "Automated human skull landmarking with 2D gabor wavelets." Physics in medicine and biology (2018).

Atilla Gül, **de Jong, Markus A.**, Jan Pieter de Gijt, Eppo B. Wolvius, Manfred Kayser, Stefan Böhringer, and Maarten J. Koudstaal. "Three-dimensional soft tissue effects of mandibular midline distraction and surgically assisted rapid maxillary expansion: an automatic stereophotogrammetry landmarking analysis." International Journal of Oral and Maxillofacial Surgery (2018).

About the author

Markus de Jong was born on August 30th 1982 in Workum, the Netherlands. There, he followed primary education at the Sint Ludgerus school, followed by secondary education at the Titus Brandsma College/Marne College in Bolsward. In 2000, he began his 5-year ir. qualifications in Computer Science at the University of Twente at the Leeuwarden department ('vestiging Friesland'). He continued his studies in Enschede, chose the track Human Media Interaction and finished with a thesis entitled "Politeness and Alignment in the Virtual Guide" under supervision of Mariët Theune, that also included a publication. In 2007, he started on forensic science courses and obtained his master's degree in 2010 with a thesis entitled "Blind Cases and Complexity in Firearms Examination" under supervision of Huub Hardy, Rob Hermsen and Wim Kerkhoff. During that time, he also started a pre-master year for psychology and obtained his master's degree for that course of study in 2010 as well with a thesis entitled "Motor skill under pressure: games, stress and automaticity" under supervision of Elger Abrahamse and Elian de Kleine. In 2011, he started his PhD at the Erasmus MC medical center that is described in this paper. Under the supervision of Eppo Wolvius, Maarten Koudstaal and Manfred Kayser and daily supervision of Stefan Böhringer in the Leiden UMC, he completed this PhD in 2019. He currently works as a post-doc on capturing bias in video news at the Vrije Universiteit Amsterdam.

PhD Portfolio

Summary of PhD training and teaching activities

Name PhD student:	M.A. de Jong	
Research School:	NIHES	
PhD period:	02-2013 - 01-2018	
Promotors:	E.B. Wolvius, M. Kayser	
Supervisor:	S. Böhringer, M. Koudstaal	
1. PhD training		
	Year(s)	**Workload**
General academic skills		
Writing grant proposal (AOCMF)	2014	20 hours
Grant progress reporting (AOCMF)	2015-2016	10 hours
Writing amendment ethical research protocol (METC)	2016	24 hours
Research skills		
Cluster computing, LUMC	2015	8 hours
In-depth courses (e.g. Research school, Medical Training)		
Computer Vision, Delft University	2013	5 EC
ASCI Front-end Vision, Eindhoven University	2014	3 EC
Advanced Genetic & Omics Data Analysis, LUMC	2015	2 EC
Presentations		
Departmental seminar presentations (LUMC)	2013-2016	75 hours
CMF Symposium Rotterdam 2014	2014	16 hours
CMF Symposium Rotterdam 2015	2015	16 hours
Facial Genetics Workshop London 2016	2016	16 hours
International conferences		
European Mathemetical Genetics Meeting 2013	2013	1 EC
Seminars and workshops		
3D face recognition symposium, Enschede	2013	8 hours
Bi-weekly Biomedical Imaging Seminar at Erasmus MC	2013-2017	86 hours
Weekly seminar at LUMC Medical Statistics dept.	2013-2017	162 hours
CMF Symposium Rotterdam 2014	2014	8 hours
Facial Genetics Workshop London 2016	2016	8 hours
CMF Symposium Rotterdam 2015	2016	16 hours
Didactic skills		
-		
Other		
-		
2. Teaching activities		
	Year(s)	**Workload**
Lecturing		
Supervising practicals and excursions		
Wetenschappelijke Vorming-1 practicals (Statistics, LUMC)	2014	16 hours
Wetenschappelijke Vorming-1 practicals (Statistics, LUMC)	2015	32 hours
Wetenschappelijke Vorming-1 practicals (Statistics, LUMC)	2016	4 hours
Supervising Master's theses		
-		
Other		
-		

Words of Gratitude

After many years, and now that my PhD has been chronicled in this dissertation, I would like to acknowledge the people that got me here.

Firstly, I would like to thank Eppo Wolvius and Manfred Kayser for being my promoters and for making this PhD possible. Together with my second co-promotor Maarten Koudstaal, I thank them for their inspiring lines of work and their fascinating research of which my PhD could be a part of.

Next, I would especially like to thank my first co-promotor and daily supervisor at the LUMC, Stefan Böhringer. His positive and ambitious attitude, his encouragements together with his broad scientific skill-set helped me throughout the years through the many ups and downs of this PhD.

I also would like to express gratitude to my committee members Hans Bosch, Ferdinand van der Heijden and Marcel Reinders, as well as Wiro Niessen, Thomas Maal and Fernando Rivadeneira. Thank you for your time and input.

Looking back, this PhD would not have been possible without previous education and my master's degrees as described in *About the author*. I am very grateful to all my teachers and supervisors during the years preparing me for my scientific career.

At the Erasmus MC Genetic Identification department, where I spent most of my time in the first years of my PhD, I would like to thank Mannis, Oscar and Fan for their company whilst sharing a room with me. Also at the Erasmus MC, I would like to thank Atilla Gül and Pieter de Gijt for their cooperation in several papers, as well as special thanks to Joan Saradin who generously allowed me to use her lab for experiments. I also would also like to thank Andreas Wollstein for laying the groundwork for the algorithm and helping me getting started.

Of course, Stefan was not alone at the LUMC, which is why I would like to thank the department of Medical Statistics for their generous hospitality and positive atmosphere, with Theo Stijen as departmental head in particular, as well as Ewout Steyerberg who succeeded him in my final years there.

Over the years, I have shared my room at the LUMC with many different PhD-students and postdocs, such as Georgios, Alexia, Jesse, Carlo, Mia, Davide, Ningning and Irene, who were all somehow able to withstand being bombarded with my stupid jokes. I've also really enjoyed the daily lunches, often a highlight of my day, with many others such as Roula, Bruna, Renaud, Dimitri, Roberta, Theodor, Angga, Ivonne, Jakub, Kate, Mar, Szymon, Said and Ramin. Thanks a lot, and I wish you, and anyone I might have forgotten to mention here, all the best!

Arthur and Lieuwe, thanks for being my good friends for many years, and thank you for being my paranymphs. It will be time for a 'biertje derbie' soon.

Of course, my prolonged studies would not have been possible without the support of my loving parents. Thank you for giving me the opportunity to do what I loved and thanks for your patience. Tige tank!

Finally, it's time to thank my family in Dordrecht. Elena, thank you for the ever loving support and for keeping me sane over the years by talking me trough many ups and downs and giving me advice and believing in me. I wish you all the best in your own PhD and I hope to return the favor by being there for you when you go through yours! Anastacia and Daniëlle, I am so happy to know you and to

spend my life with you and Elena as a family. I love you all!

Acknowledgments

This PhD was supported by a grant from AOCMF (AOCMF-13-12K).

The TwinsUK study was funded by the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013). The TwinsUK study also receives support from the National Institute for Health Research (NIHR) BioResource Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London.